Hardware-Software Co-development for Audio and Video Data Acquisition and Analysis

PhD Thesis

György Kalmár
Supervisor: László G. Nyúl, PhD

Doctoral School of Computer Science
Department of Image Processing and Computer Graphics
Faculty of Science and Informatics
University of Szeged

Szeged
2020

”Scientists study the world as it is,
engineers create the world that never has been.”

(Kármán Tódor)

Contents

Chapter 1

Introduction

Pattern recognition is the science of algorithmic identification of patterns in data. The objective is usually classification, which requires the assignment of pre-defined labels to data samples. This may require the implementation of a training process when labeled data points are employed to discover the inner-class regularities between 'similar' samples. The data may be diverse and pattern recognition includes many different fields and applications, such as signal processing, image analysis, or speech recognition, to name a few.

In the past, pattern recognition involved algorithms designed by experts. Later, machine learning procedures emerged to solve the classification section of the problem. These methods were trained on compressed and irredundant input, called feature vectors. These features were hand-crafted and required prior knowledge related to the specific areas. Today, with the advent of artificial intelligence (AI), pattern recognition is dominated by AI algorithms, and the term itself usually implies that some sort of machine learning method is involved. These AI procedures take the training dataset, perform automated feature extraction and selection, on which the classification task is optimized. This scenario is called supervised learning, which requires a (potentially large) labeled dataset during the training process. Its counterpart, unsupervised learning, such as clustering, processes training data that has not been hand-labeled, and attempts to find inherent patterns in the data that can later be adopted to determine the correct label for new data instances. In the current work, only supervised methods are discussed ranging from classical solutions to modern, AI-based algorithms.

The key component of any data analysis procedure is data it is trained and tested on. Mainly, quality and quantity are significant. To explain this, the reader can visualize, for example, the classification problem as the separation of an **N**-dimensional space into **M** disjoint subspaces, where **N** is the length (dimension) of the data (signal, image, audio, text, feature vector, etc.) and **M** is the number of classes. The labeled data points fill discrete points of this space. The task is to separate the labeled

point-clouds in space so that the classification problem achieves maximal accuracy. In this case, it can be seen that as long as the points belonging to the distinct classes represent these classes well, the problem can be solved properly.

Classical solutions are heavily based on humans. The human mind and imagination are able to generalize in an extraordinary way that even a small amount of data is enough for an expert with field-related knowledge to create a well-generalized and accurate data analysis procedure. Back to the previous example, this means that the data points are spread only within a small region of the space or do not represent the classes well, but with the help of prior knowledge, the formed decision surfaces are accurate. By the time, the complexity of tasks increased in parallel with the rapid rise of computational power and storage capacity. The human brain could no longer mold its understandings into program code form. Data-driven procedures have come to the fore, which grab the space filled with labeled data points, as many as possible, and try to find optimal decision surfaces. These methods are ”data-sensitive”, meaning that the number of data points, their appropriate distribution, and noisiness are crucial and fundamentally limiting the accuracy of the results. As good the data is, as accurate the methods can get.

Modern pattern recognition and data analysis methods, like neural networks (NNs), usually run on Graphical Processor Units (GPUs) that can execute billions of floating-point operations in a single second. These modern tasks are usually high-level, meaning that complex analysis of huge amounts of structured data is required to produce high-level descriptors. These tasks include, for example image segmentation, video analysis, object recognition, speech recognition, natural language processing, and many more. These fields form the backbone of modern data science. Contrary, they provide only a moderate portion of applied pattern recognition as the delegation of such methods to end-user products has started slowly in recent years, which market is one of the main drivers of pattern recognition advancements. Despite the fact that decision-making, called inference, is a lot cheaper in terms of computations than training, running a complex neural network on embedded devices is still challenging. These systems have restricted resources like power, memory capacity, clock frequency, size, etc. and real-time responses are usually required. Therefore, only simple and well-optimized methods can be run to process the data-flows from the sensors in real-time. In these fields, classical and hybrid pattern recognition algorithms are still relevant.

This work presents three data analysis and pattern recognition applications ranging from low-level audio classifier embedded systems to high-level image segmentation algorithms. A common approach connects them that is with hardware-oriented modifications and related software co-development the corresponding tasks became feasible, easier, or more accurate.

Hardware changes are complicated in data analysis applications and they usually imply the modification of the software as well. Many times, sophisticated algorithms can provide solutions to hardware-related drawbacks, but spending processing capacity to correct these anomalies is a waste of time and energy — two such quantities that are limited, for example, in embedded systems. Furthermore, if these anomalies impact data quality, even the high-level data analysis algorithms are affected, as was explained earlier. Many times applications reach a point, where further software optimizations offer limited improvements compared to the energy and time required to implement them. Noisy, low-quality data can be enhanced but with restrictions. In these cases, even simple hardware changes or extensions may lead to improved data quality or new analysis directions. This consideration is applied in the current PhD. work.

One such example of this problem is presented in Chapter 2., where an animal-borne acoustic gunshot detector is introduced. The main idea is to put a single wearable device on elephants that detects gunshots near to the animals and sends GPS-tagged alerts in case of poaching activity. Gunshot detection requires the constant monitoring of the environment with a relatively high sampling frequency. Therefore, the power consumption of such systems is high, which limits their applicability in wearable designs. Another problem is that with such restricted embedded devices only simple detection algorithms can be run that may lead to very costly false-positive alerts. To solve these problems, hardware-level modifications were carried out. With an acoustic delay line structure, the power consumption was reduced significantly and the detection accuracy increased at the same time. The developed system was tested in real-world conditions.

In Chapter 3., another sound classification problem is presented. It is well-known that loudspeakers are capable of recording sound. This microphone-like feature is investigated from theoretical and practical perspectives. The idea is to extend the hardware of loudspeakers with a device listening on their driving lines so that acoustic event detection becomes feasible. The original functionality of the speakers is retained, but when it is not exploited, a new function becomes available and it requires a simple hardware modification only.

Chapter 4. presents an image processing application that aims to detect and measure the pupil regions of rats in pupillometry videos. The algorithm supports a medical research, which examines schizophrenia-like symptoms in rats. The chapter presents the development process of the experiments. In the first phase of the research, the videos' quality was poor and a complex, classical image processing method was required to handle the drawbacks. To enhance the video robustness and quality, the recording camera experimentation hardware was extended. With the revised experiments, the resulting videos improved in quality so that data-driven methods became feasible. A Convolutional Neural Network (CNN) was trained to

perform the pupil segmentation task, which provided a new solution to the pupil measurement problem. The key point in this chapter was the hardware extension of the recording camera.

The above-mentioned chapters present different data analysis applications, however, many similarities can be mentioned. The same approach appears in them, which is the hardware-oriented modifications and corresponding software developments that led to improved solutions. Each chapter also includes traditional and also modern, machine learning-based algorithms and compares them. The order of the chapters is also important. As the chapters progress, the hardware capabilities and the complexity of the problems increase.

1.1 Contributions

The ideas, figures, tables and results included in this thesis were published in scientific papers (listed at the end of the thesis). In a nutshell, the author is responsible for the following contributions:

Chapter 2.: The idea of the delay line structure and the two-phase wake-up procedure. The development of a multi-stage gunshot detector that utilizes effectively the extra information originated from the novel wake-up method. Investigation of possible AI-based gunshot detector methods.

Chapter 3.: The theoretical and practical analysis of loudspeakers in microphone mode (referred to as reverse mode) operation. Utilization of this capability by developing an embedded device that can detect suspicious events on the speakers' driving lines.

Chapter 4.: Automated and improved pupillometry of rats to support related medical research. Designed a classical image processing algorithm for low quality videos. Improved the image quality by extending the experimentation hardware, which led to an AI-based pupil segmentation algorithm.

Chapter 2

Animal-Borne Anti-Poaching System

Acoustic gunshot detection has been an area of active research in recent decades. Many algorithms and systems were developed to realize a low-cost, low-power, and reliable solution. In this chapter, a wearable, animal-borne gunshot detector is introduced, which offers ultra-low power consumption and enhanced detection accuracy. The device is integrated into consumer GPS tracking collars, therefore, the combined system is able to send GPS-tagged gunshot alerts to support law enforcement.

The main novelties of the work are a specially designed wake-up procedure that allows low-power constant listening and a two-domain based gunshot detection algorithm. Hardware-level modifications solved the critical power-consumption problem and provided more information for the detector, which resulted in enhanced classification accuracy.

The structure of the chapter is as follows. Section 2.1 and Section 2.2 introduce the problem and list related research and notions. The succeeding sections explain our approach and system design, and detail the novel wake-up mechanism and compare it to traditional solutions. In Sections 2.7 and 2.8, the architecture of the developed system is presented. In Section 2.9, the evaluation of the system and its results are included. Section 2.10 presents a brief exploration to possible data-driven approaches. In the closing section, the chapter is summarized and final thoughts are aggregated.

2.1 Introduction

Poaching is one of the primary drivers of wildlife decline, notoriously being listed among the top five drivers of biodiversity loss [11]. While it targets a vast array of animals, the highest valued are increasingly large-bodied, charismatic species that are particularly susceptible to overharvest due to their slow rate of population growth. The impact extends beyond the demise of the targeted species. Poaching-caused wildlife declines can have serious implications for ecosystems, where the removal of

animals from illegal harvest can have cascading effects on other species and even productivity of the system as a whole [28].

Poaching is, by definition, illegal. As such, interventions to reduce it typically follow classic law enforcement approaches. However, wildlife poaching tends to occur in remote areas, with low human densities, where detection is difficult. In addition, poaching of large, high-value species is militarized and can be driven by global crime syndicates. As such, local wildlife agents can be operationally overwhelmed, not only in terms of law enforcement equipment, but often due to the limited capacity to monitor widely distributed animals. The development of technologies designed to overcome the challenges of remote wildlife protection that can enhance protective efficacy is needed.

Animal-borne sensors, particularly GPS-equipped collars, are used to enhance real-time wildlife protection [90]. Tracking technology offers near real-time access to the location of animals and sensor data collected on the collar. In addition, GPS tracking has become a key approach in wildlife conservation efforts focused on resolving broad landscape management issues [93]. As a result, the application of radio collars on species is becoming one of the most common tools for wildlife monitoring in ecology and conservation. Innovations that can be integrated into tracking systems can immediately scale, offering broad application. GPS tracking is currently integrated with many anti-poaching systems, but primarily as a means to deploy assets in the vicinity of at risk individuals, though interest in using movement data to identify exposure to risk is increasing [90]. While these data streams have been valuable to resolve a number of conservation challenges, these systems have not been particularly effective in identifying poaching in real-time [76]. Detecting poaching events is a critical need to provide actionable information for law enforcement.

The chapter presents a novel approach to anti-poaching: a ballistic shockwave detector integrated into an existing GPS collar to provide real-time alert of shots fired near the protected animal. Note that the objective is not to prevent the current poaching, but to notify the authorities so that they can apprehend the perpetrators and prevent future attempts. The work focuses on the two main technical challenges: power consumption and gunshot classification accuracy. The device needs to last two years on a single charge while continuously listening for shots. Note that no current energy harvesting technology is applicable due to the rough conditions. Law enforcement response in remote areas are extremely resource intensive, so false detections must be kept at an absolute minimum.

2.2 Related work

The use of sensors to identify security risks to animals represents a key opportunity that can advance wildlife protection, particularly in remote areas where standard

anti-poaching techniques are challenged. Currently, the use of on-animal tracking systems for identifying mortality events has largely relied on beacon based signals which identify immobility after an extended period (usually 24 hours) [86] or clustering algorithms of locations to identify unnatural immobility. Once these indicators have been signaled and detected, the location can be investigated to assess if a poaching event happened. However, animals may sleep for hours and the transmission of GPS data can be delayed depending on the schedule of the collar. The delay offered by clustering or beacon based identification of immobility often means the detection of the security event is outside of an effective operational window. This limits their utility for direct intervention.

Increasing interest in sensors designed to detect immediate events is driving the development of multiple add-ons to tracking systems [86, 92]. These sensor-based approaches offer unique insight into behaviors of interest but tend to have limitations that inhibit their utility for anti-poaching solutions. Biophysical monitors are promising, but often require invasive approaches (e.g., surgical implants) and have limited lifespans. More commonly, activity sensors, such as accelerometers, are being employed to provide fine-scale information on the status and activity of a tagged individual [72, 81, 92]. However, the use of accelerometer data as a means to detect mortality has proven difficult, given that behavioral identification using accelerometers is prone to false positives.

2.2.1 Gunshot detection

There are two acoustic events associated with firing a typical rifle. The muzzle blast originates at the gun itself and spreads spherically at the speed of sound. It is the result of the propellant of the ammunition exploding inside the barrel of the gun. The second event is called the ballistics shockwave and it is caused by the bullet travelling faster than the speed of sound. This sonic boom creates a conical waveform whose tip is the bullet and that expands at the speed of sound. Both of these events can be picked up by a microphone.

A muzzle blast is a high energy event characterized by a rapid rise. However, the time-domain signal shape depends on the rifle and ammunition used and is greatly affected by the environment due to echoes. Also, the source of the muzzle blast is the gun, which can be quite far from the animal. Finally, the sound energy and, hence, the detection range of the muzzle blast can be significantly lowered by a suppressor. For all these reasons, the muzzle blast is not an ideal signal for an anti-poaching sensor.

In contrast, the ballistic shockwave is a unique acoustic phenomenon (Figure 2.1). Its shape in the time-domain resembles a capital N with sub-microsecond rise time and a total signal length of a few hundred microseconds depending on the caliber,

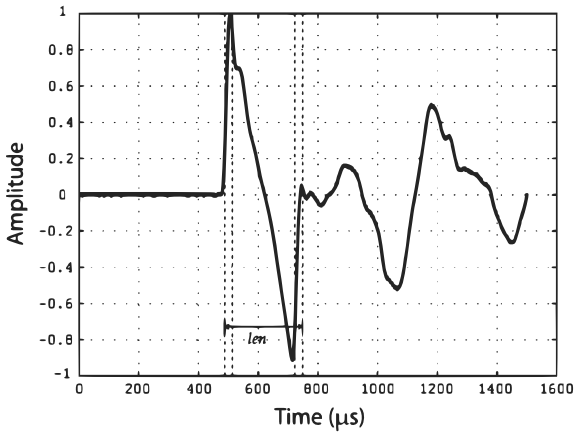


Figure 2.1: *Shockwave of an M16 projectile*

speed, and miss distance, that is, the minimal distance between the sensor and the trajectory of the supersonic projectile [91]. It is also a high-energy event, especially at a short miss distance. As such, it requires microphones with low sensitivity, but with a superb frequency response. The further the microphone is from the trajectory, the more the N-shape of the signal gets distorted by the air acting as a low pass filter. As such, the effective and reliable detection range of the shockwave is about 50 meters. However, the miss distance will be small for poachers that are shooting at an animal and, therefore, the sensor. Given that the source of the acoustic event, the projectile, will be closer to the sensor than the gun, the samples recording the shockwave will precede those of the muzzle blast when arriving at the microphone. Finally, only subsonic rifles can produce projectiles that do not generate a shockwave, but their effective range is much shorter than those of regular rifles making them less commonly used for poaching. Two widely used poaching rifles are the AK-47 and the M16 (or its civilian variant, the AR15) due to their proliferation around the world. Both are supersonic, as are almost all big game hunting rifles. All of these characteristics make the shockwave the ideal target signal for an animal-borne gunshot detector.

2.2.2 Acoustic anti-poaching systems

There has been at least one attempt to employ a commercial gunshot detection system for anti-poaching. ShotSpotter [78] has been in use in various U.S. cities to alert police to the location of gunshots. In 2014, a small system of a few sensors was tested in Kruger National Park to detect and localize muzzle blasts related to rhino poaching. The sensors were deployed at fixed locations covering a few square miles. While there is no published evaluation of the experiment in the scientific literature, it is notable that this system is not currently employed. It is probable that unreliable

acoustic classification due to attenuation and distortion limit the utility of this system in large wilderness areas.

There are acoustic wildlife monitoring systems in use across the world, which are often deployed for extended periods. Some of these are able to detect gunshots [56, 94]. Given the few hundred-meter effective detection range for muzzle blasts, and the fact that one needs three sensors to locate a point source, these systems can only protect very small geographic areas. Covering Kruger or Serengeti would require millions of sensors, which is not practical.

2.3 Approach

Current acoustic shot detectors aimed at identifying poaching events are inadequate, given they are statically deployed and rely on muzzle blasts resulting in a limited detection area. Hence, these systems do not scale geographically. An animal-borne shot detector, on the other hand, protects the animal and its herd, not an area. Furthermore, given the animal is the target, the ballistic shockwave can serve as the primary event making classification more accurate.

The aim of this work is to integrate an acoustic shockwave detector into existing GPS tracking collars. Current tracking systems tend to record the GPS positions hourly and upload those positions several times a day. However, the presented system can be event driven, such that when a gunshot is detected, it immediately sends an alert with the last recorded GPS location, then records and sends a new position. This ensures event identification should poachers destroy the unit before the recording and sending of a new GPS position. Note, that the first shot will always be aimed at the animal and not the sensor, meaning that the system cannot necessarily save the animal wearing it. Instead, the goal is for law enforcement to be able to apprehend the perpetrators and hence, prevent further poaching. It can also act as a deterrent if the poachers realize the increased risk of getting caught.

Our initial target species is elephants because they are subject to high levels of poaching across Africa and Asia, they are being tracked in numerous locations, and their large size provides the fewest mechanical constraints in terms of size, hence, batteries. However, sensors are difficult to deploy on elephants and they are rough on the collars. Consequently, the units need to last multiple years on a single charge and the enclosure needs to be very robust. In addition, false positives are a serious concern given the remote locations where elephants roam make responding to an alert resource-intensive. Hence, the false alarm rate of gunshot classification must be kept to a minimum.

Existing wearable gunshot detectors for the military only last a day or less on a single charge [68, 88]. This is because they continuously sample and process the recorded acoustic signal. They typically use multiple microphones and high sam-

pling rate for Angle of Arrival estimation. But even a single microphone and lower sampling rates would result in an order of magnitude lower power consumption at best. A larger battery can gain another $10\times$ improvement. The single greatest technical challenge for the anti-poaching sensor is the requirement for another order of magnitude improvement in battery life. Due to the harsh conditions and limited size, solar or mechanical energy harvesting cannot address such energy requirements. Our solution employs an ultra-low-power microphone attached to the wall of the protecting box that wakes up the rest of the system using an analog threshold trigger. The acoustic sound wave is guided through a hole and a thin tube — a kind of acoustic delay line — to an electret microphone, delaying it just enough so that the entire event can be captured without information loss.

The second significant challenge is the need to virtually eliminate false positive detections. Having two microphones with different characteristics is useful. But we also add an accelerometer so that possible gunshots can be correlated with changes in the motion of the animal. Fall, immobility, or a panic run after a shot candidate will increase the confidence in detection accuracy. To validate the design of the sensor and finalize the detection algorithm, the first prototype units have SD cards on-board that store 3-axis accelerometer data continuously as well as all detected acoustic events. Due to the added power consumption of data storage and in order to progress with the development at a reasonable rate, the initial deployments are expected to last only a few months. Based on the data gathered, the hardware and software of the system will be revised and finalized.

2.4 Tracking collars

A popular sensor model—made by Savannah Tracking [77]— has been selected as the initial platform for integration with the gunshot detector. It is already widely used on elephants and, with different collar designs, on other species in many areas of Africa and Asia. The collar itself is made from a 135 mm wide 10 mm thick cotton fibre and rubber transmission belting. The electronic board and battery are housed inside a half-moon shaped nylon casing placed on the top of the collar in a stainless-steel metal housing. Note that once ready for deployment, the sensor enclosure is filled with resin to protect the electronics from the elements. To keep this unit on top of the neck, a steel counterweight is placed under the neck which further acts as the place for connecting the belting during deployment (Figure 2.2). The total collar weight is 14 kg.

Position acquisition is done using GPS localization. The intervals between the position recordings and the lengths of averaging periods can be defined by the user to balance between the accuracy and power consumption. Positional data is transmitted at user-defined intervals, typically every 3-6 hours, to a cloud-based server

Figure 2.2: *Deployed tracking collar*

via the Iridium satellite SBD service [44]. Optionally, a GSM modem can replace the satellite-based communication for deployments where cell coverage is reasonable. Up to 24 positions can be sent in a single message, hence up to hourly positions can be included in a single daily report. The communication is half-duplex, allowing the reconfiguration of parameters of the data collection schedule after collar deployment. Data access, collar reconfiguration, database management, visualization, and animated replaying of the data are all provided via a cloud-based server.

The tracking unit contains 4 lithium D-cell batteries providing a total of $230\,\mathrm{Wh}$ of power. With a typical schedule of 24 GPS positions and 4 data reports per day, this will provide around 10 years of lifetime, well above the expected average physical lifespan of 2-5 years. The tracking module also contains a tri-axis accelerometer collecting data at user-defined settings of $1-100\,\mathrm{Hz}$ and between $2-8\,\mathrm{g}$ sensitivity. This data is not transmitted but evaluated on-board in real-time. Activity patterns suggesting unusual behaviour (multiple excessive motions spikes) or mortality (immobility) will trigger an alarm response, which contains a GPS location and the specific type of alarm triggered. Once received by the server, this alarm is forwarded via e-mail and/or text message to a collar-specific list of contacts. Hence, shortly after an animal has started behaving in an unusual way, users will be alerted automatically with a message containing the animal's current position and the type of alarm triggered. While the accelerometer data alone is prone to false positives, combining gunshot detection with motion sensing may result in a real-time and more robust poaching detection system.

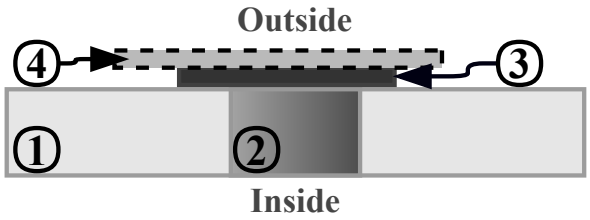


Figure 2.3: *Structure of the acoustically permeable waterproof mechanical protection. (1) metal wall, (2) small hole, (3) waterproof acoustic vent and (4) metal mesh. The vent and the metal mesh are held in place by a metal plate bolted to the wall with a hole in the middle (not shown).*

2.5 Mechanical protection

Given the strength of elephants, the mechanical protection of an acoustic sensor is challenging. The protective material needs to be strong and thick to survive the required lifetime of the sensor. The steel box, thick nylon and the resin filling offer reliable protection but also reduce the acoustic energy and the signal-to-noise ratio inside. Another undesirable effect of the rigid metal wall is the greater attenuation at higher frequencies in such dense medium [49]. The unique aspect of a shockwave is its extremely rapid rise time, which is heavily distorted by the enclosure.

To reduce these unwanted effects, a small hole is drilled into the metal wall to enable unattenuated sound propagation into the box. With this solution, the sound quality is preserved, but the waterproofness is lost. Fortunately, the same problem arises in today's handheld devices and acoustic waterproof vents with favorable sound transmission properties are readily available. These highly breathable expanded polytetrafluoroethylene membranes vibrate easily, rapidly equalize pressure, and offer protection up to IP68 [34]. The sound transmission loss is below 2dB, which is negligible. These vents are used to cover the drilled hole, but they have a sensitive mechanical structure that needs additional protection when applied at the external surface of the box. Therefore a metal mesh with strong mechanical properties and without sound-distortion effects is used to protect the hole and the acoustically transparent venting. The final structure of the mechanical protection is shown in Figure 2.3.

2.6 Wake-up mechanism

Acoustic gunshot detection is a pattern recognition task that requires the constant recording and processing of environmental sounds. In wearable devices, there are trade-offs between power consumption, sensitivity and information loss. In this section, a novel wake-up mechanism is presented and compared to the state-of-the-art.

2.6.1 Traditional solutions

In many applications, *almost* constant listening can be achieved by using *analog threshold-based* wake-up circuitry to trigger recording a very short time after the acoustic event has started. Usually, the initial loss of information is negligible compared to the full length of the pattern of interest. With specialized microphones and architectures, the initial wake-up delay can be as low as 100 μs [89], while keeping the power consumption very low.

In Section 2.2, we explained the basics of gunshot acoustics and showed that if the listener hears the shockwave, then it must be the initial impulse-like section that reaches the microphone first. Preserving the quality, and ensuring the recording of this part of the shockwave is crucial to maximize classification accuracy. By using the previously mentioned wake-up mechanism, up to $1/3$ of the N-wave pattern would be lost due the 100 μs initial delay. The classical analog wake-up mechanism, therefore, cannot be used in shockwave-based gunshot detection systems.

To solve the information loss problem, we have to take into consideration another classical solution, the *always-on listening* method. The main idea is to turn the microphone on and constantly keep pushing digitized samples into a circular buffer to maintain a short signal history in the memory. When a gunshot event happens, samples collected before the shockwave arrival are already stored in the memory enabling the recording of the full N-wave. However, this solution requires active clock sources for the microcontroller and for its peripherals. By optimizing a system for this solution, low power consumption can be achieved, but it is still significantly higher than the previously described analog mechanism.

2.6.2 Delay line wake-up mechanism

The combination of the ultra-low power consumption and the preservation of a full shockwave is essential in wearable gunshot sensing. To satisfy both requirements, a two-domain based wake-up mechanism was introduced. The proposed structure is based on a kind of acoustic delay line enabling a two-phase wake-up procedure, illustrated in Figure 2.4. This solution uses two types of microphones: a contact and a traditional electret microphone. The contact microphone, or pickup, is a transducer that converts the vibration of the surface it is mounted on to voltage by utilizing the piezoelectric effect. In our case, this microphone is attached to the internal side of the metal sensor enclosure. The second, traditional microphone is placed at the end of a 3.5 cm-long tube. This tube serves as a waveguide and soundproofing for the incoming sound waves that enter the box through the protected hole.

The main idea behind this structure is to wake up the data acquisition system from deep sleep mode when the acoustic waves reach the metal wall and delay the sound waves by the tube to ensure the required amount of time for the system to

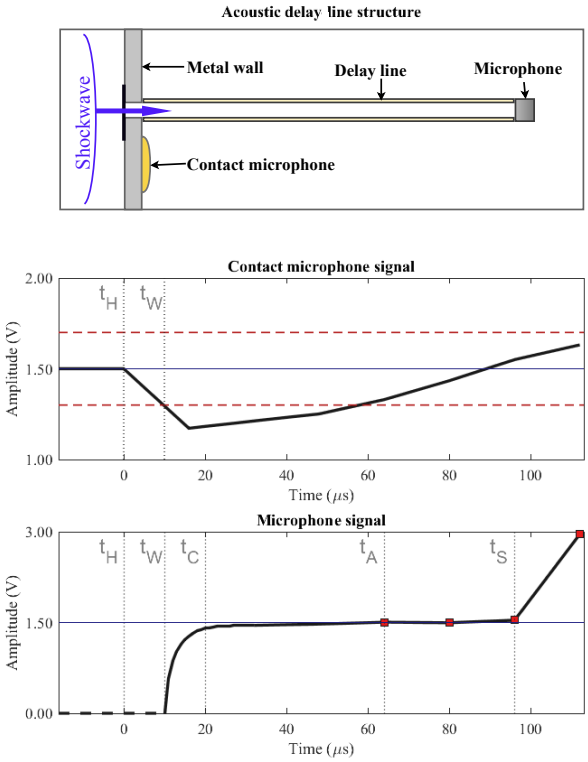


Figure 2.4: *Details of the wake-up mechanism. In the top row the structure of the delay line and the propagation of a shockwave from left to right are shown. Below that, the microphones' signals synchronized to the shockwave's progress are plotted. Different time points of the event are marked:* t_H *- the shockwave hits the wall;* t_W *- wake-up signals are generated by the contact microphone;* t_C *- CPU and microphone switch to active mode;* t_A *- the first ADC sample is collected;* t_S *- the shockwave reaches the electret microphone.*

prepare for data collection. This is possible since the speed of sound is negligible compared to the speed of light and the voltage generated by the contact microphone travels at the latter.

Timing of the wake-up procedure

The time required by the sound to travel through the tube can be calculated from the length of the tube (l) and from the speed of sound in air (c). Latter varies by the temperature, and in Africa, extreme hot weather is possible. Using reasonable limits, the propagation time, T_{prop} becomes:

$$T_{prop} = \frac{l}{c} = \frac{0.035\,m}{345 \pm 20\,m/s} = 101 \pm 7\,\mu s.$$

In Figure 2.4, the timing of the wake-up mechanism is presented. The shockwave arrives from the left side and propagates through the tube to reach the microphone at the end. At time point t_H, the shockwave hits the metal wall and pressure waves convert to vibration, thus voltage is being generated by the contact microphone. When the voltage level crosses the previously set threshold level, at t_W, an analog comparator wakes up the microcontroller and turns the microphone's power supply on. Our MCU needs 10 μs to wake up from deep sleep mode, so at t_C the CPU is active and enables the analog-to-digital converter (*ADC*). Approximately 60 μs has passed and at t_A the ADC has already collected the first digitized sample from the stabilized microphone signal. Around 40 μs later, at $t_S = t_H + T_{prop}$, the shockwave reaches the electret microphone and the leading edge is captured.

Advantages of the proposed mechanism

In the delay line structure, the power consumption before a wake-up event is minimal, as only the contact microphone is active, which does not consume energy, instead, generates voltage. Additional elements like an amplifier and comparators are needed, but ultra-low-power parts are available. The gunshot detection system spends most of the time in deep sleep mode and even the electret microphone is turned off. This property offers a long lifetime to the detection system.

In addition, the delay line also enables data acquisition without information loss. This has a big positive impact on the detection accuracy as it will be presented in Section 2.9. When the system starts sampling, it perceives a short section of the

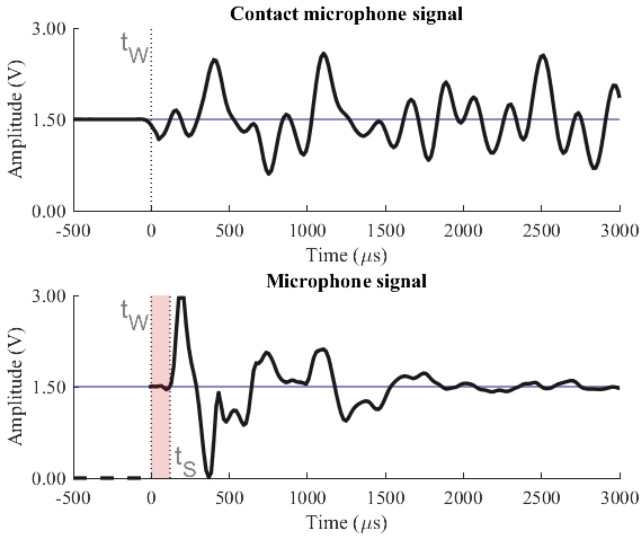


Figure 2.5: *Gunshot recorded with the acoustic delay line wake-up mechanism. The shockwave woke up the system at time point t_W, and reached the microphone at t_S. The shaded delay region between the two events is provided by the acoustic delay line.*

signal's history, which happened in the past, before the wake-up event.

A gunshot event recorded with the delay line structure is presented in Figure 2.5. The shaded region marks the delay between the wake-up event t_W and the arrival of the shockwave at the microphone t_S. This delay period between the two events is minimal when the position of the shockwave impact point coincides with the location of the drilled hole in the wall, that is, when the projectile trajectory is on the same side of the box as the hole. If the shockwave hits the box from other directions, the delay becomes longer, since the speed of sound in steel is over $10\times$ higher than in air, causing the vibration to reach the contact microphone almost immediately, while the sound propagating in the air needs more time to get to the hole around the enclosureevent.

Note that using only the low-power contact microphone alone is not an option since it would still miss the beginning of the shockwave. Moreover, the vibration of the metal wall that it measures does not preserve the characteristics of the shockwave as can be seen in Figure 2.5. Mechanical impacts on the sensor enclosure generate similar signals. Also, the two signals with different characteristics are of great help in gunshot classification. See Section 2.8.

Comparison with active listening

The fast processing of the recorded signals requires significant amount of energy in any solutions. However, it is completed in only a few hundreds of millisecond in the case of short, impulsive gunshot events. Therefore, even at high event frequencies like 1 event/minute, the total consumption is dominated by the low-power listening mode.

During the listening period, the delay line wake-up structure only runs an amplifier and two analog comparators as the main energy consuming parts. In contrast, the active listening method constantly requires enabled clock sources, active MCU and analog-to-digital converter, turned-on microphone and amplifier. Both approaches were implemented with the same ultra-low-power hardware components, and the corresponding power consumption values were measured in the idle listening phase. The delay line structure needed $102\,\mu A$, and the active listening method required $832\,\mu A$. The detailed comparison of the two mechanisms' consumptions is illustrated in Figure 2.6.

The significantly lower power consumption of our approach offers $8\times$ *longer lifetime*, or the same lifespan with batteries having significantly smaller size. A lighter and more compact sensor makes the approach feasible for smaller animals too.

A potential drawback is some distortion of the acoustic signal caused by the delay line tube. This effect was analyzed in the frequency domain by using chirp excitation signals. In these harmonic cases, we experienced some distortions at multiple frequencies due to resonance and standing waves in the tube. However, with real-world

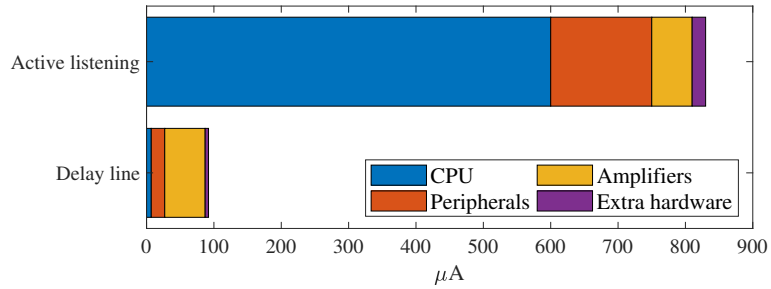


Figure 2.6: *Power consumption comparison of the active listening and delay line wake-up methods.*

gunshot tests and using an external reference microphone as a baseline, we didn't experience noticeable impact on the shockwave signal shape.

The idea of using an acoustic delay line does not necessarily imply the utilization of a contact microphone. In our case, it ensures even lower power consumption and enhanced detection accuracy (explained in Section 2.8), but other applications may use different structures. For example, using an ultra-low-power traditional microphone as the wake-up source instead of the contact microphone is also possible. Furthermore, longer delay lines allow access to a longer history of the signal prior to the wake-up event, so less impulse-like patterns can be recorded too.

2.7 System architecture

In this section, the hardware architecture and the most important software components are presented. A small, low-power sensor board was designed to capture, process and optionally store the acoustic signals. The developed software controls the timing of the wake-up mechanism and performs gunshot classification.

The low-power gunshot detector subsystem presented here is integrated with an existing tracking collar. The connection between the two systems is a simple two-wire protocol that can transmit alerts along with a few parameters such as confidence level. The reason for the simplest possible interface is to minimize any hardware or software modifications of the existing collar system. Once a collar is deployed, it is very difficult and resource-intensive to retrieve. Furthermore, tranquilizing an animal is a traumatic experience, so it must be avoided unless absolutely necessary. Neither is over the air firmware upgrade possible due to the low bandwidth communication channel. Once deployed, the collar must work. Therefore, any modifications carry significant risks. The currently utilized collar already has an interface for plug-and-play extension with additional sensors providing alerts. Therefore, it did not require any modifications to add the gunshot detector.

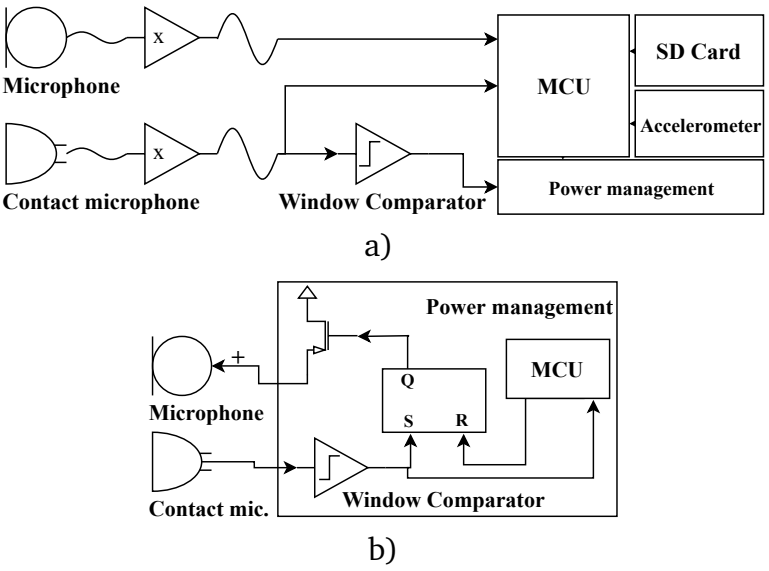
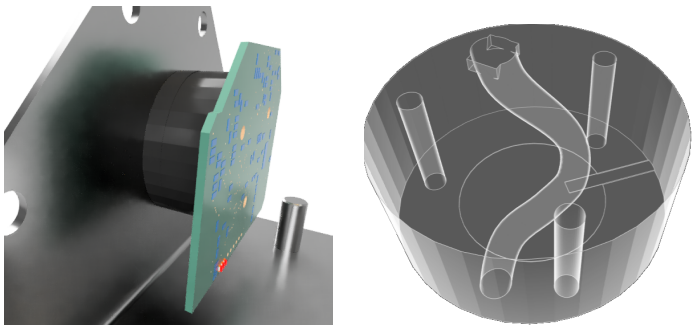


Figure 2.7: *Hardware components of the developed board: a) abstract hardware structure, and b) details of the power management subsystem.*

2.7.1 Hardware components

To implement the low-power acoustic delay line based monitoring approach, a sensor board has been developed. Figure 2.7a summarizes the main components of the system. As was explained earlier, the wake-up mechanism utilizes two microphones. Each microphone uses an amplifier for signal conditioning (biasing and amplification). The contact microphone's signal is connected to the MCU and to the power management subsystem, shown separately in Figure 2.7b. When the contact microphone signal leaves the interval defined by the window comparator, a logic level wake-up signal is generated. This signal is connected to the MCU, to wake it up through an interrupt, and to an SR-latch that stores the state of the power manager. The SR-latch controls a high-side switch, which, in active state, turns the electret microphone on. With this solution, the wake-up process is faster, because the MCU and the microphone are turned on at the same time. When the acoustic event recording phase is over, the MCU can reset the power manager's state (inactive microphone). Based on the detector output, the signal buffer can be saved on an SD card in the form of an audio file. The sensor board contains an accelerometer, which can be used to the monitoring the animal's movements after a potential gunshot detection. A panic-run, fall or total absence of motion can reinforce the previously sent alert.

An interesting problem occurred with the fast wake-up of the microphone. Traditional analog microphone signal conditioning circuits utilize capacitors for DC component removal before the biasing and amplification step. This capacitor needs to be



a) b)

Figure 2.8: *Illustration of the gunshot detector subsystem assembly (a). The sensor board is attached to the side of the metal wall with the help of a holder element. The curved tube inside the holder guides the sound to the microphone delaying it just enough to wake up the sensor board. The inner structure of this element can be seen in (b).*

charged before normal functionality is provided. The charging time of the capacitor limits the stabilization time of the microphone signal as the charging current is limited by the impedance of the feedback resistors in the amplification phase. To overcome this limitation, an active boosting circuit was used that opens a low impedance route to the capacitor in the first 40 μs of the wake-up procedure. During this period, the capacitor is being charged quickly.

The size of the existing protective box is limited and our sensor board needs to be attached to the wall. Therefore, a compact microphone and sensor board holder unit was designed, which can be 3D printed from semi-soft rubbery material instead of hard plastic. This material helps reduce the effect of the vibrations on the electret microphone. The cylinder-shaped holder has three functions. The most important one is the embedded acoustic waveguide. A 3.5 cm-long, 3 mm-wide tube, without any sharp turns or edges is meandering inside the holder, connecting the hole in the metal wall with the microphone. The curved tube design reduces the size of the holder to about half of the 3.5 cm tube length. The microphone is soldered directly onto the sensor board, which is fastened to the holder in such a way that the microphone penetrates the entrance of the tube. The other side of the cylinder presses the contact microphone to the metal wall. The entire structure is attached to the box with machine screws that are sealed for waterproofing. See Figure 2.8.

2.7.2 Software components

The two most important tasks of the sensor are rapid wake-up and reliable gunshot detection. Once deployed, the system cannot be restarted or updated and must provide continuous listening for multiple years. Reliability, real-time response and power-awareness are common criteria in embedded systems, so well-known meth-

ods exist to support the development process.

The sensor board contains an STM32 Cortex M4-based microcontroller [83]. The peripherals were configured and the initialization code was generated by the STM32CubeMX software [82]. Only hardware abstraction layer (*STM32Cube HAL*) functions were used, which offer enhanced code reliability with acceptable run-time overhead.

The developed firmware is an event-driven application that controls the wake-up logic, records the two microphones' signals at 66 KSPS sampling rate, runs the processing algorithm and sends alerts if needed. Some extra functionalities were added to the initial prototype: recording the acoustic events and continuous streaming of 12.5 Hz accelerometer data, both to the SD card.

2.8 Gunshot detection

When an acoustic event happens, the system starts recording it within 100 μs and a fast, nearly real-time decision is needed, because the risk of sensor damage is very high as poachers may try to destroy the device. Therefore, processing time must be limited mandating the use of simple algorithms. The resource-constrained embedded platform also points in the same direction.

In Section 2.2, it was mentioned that a common rifle used by poachers is the AK-47, which has a 600 rounds/minute nominal rate of fire. It means that in every 100 ms a gunshot event can happen. The supersonic speed of the projectile causes a time separation between the arrival of the shockwave and the muzzle blast to the sensor. This period can easily exceed 100 ms from reasonable ranges and it may cause an overlap of the shockwave of the second and subsequent shots and muzzle blasts at the listener's position. However, the first shockwave is almost always the first to arrive with no overlapping muzzle blast. (The only exception is when we are shooting away from the sensor and the source of both the shockwave and the muzzle blast becomes the gun itself.) We chose to record a 120 ms-long period and tried to simplify the recognition algorithm to minimize the processing time.

Note that signals are only analyzed in the time domain. Resource-intensive methods based on correlation or more sophisticated techniques [1, 32, 57, 73] would be not feasible on our constrained platform because of time, memory and energy limits. Other shot detector systems used similar design decisions in the past [79].

2.8.1 Structure of the detector

Our gunshot detector has three stages. The first stage runs in real-time, and its main functionality is to filter out false wake-up events. If the microphone samples collected in the first 3 ms after the wake-up are all below a threshold value, the

system stops recording and goes back to deep sleep mode. Below this threshold value, the recorded amplitudes are so small that no reliable detection is possible.

The second stage implements a cross-domain filtering and only runs offline. The main functionality is to filter out acoustic events caused by mechanical impacts on the box. Basically two types of events are possible. The first type is produced by sound pressure waves; the second is generated by mechanical impacts. The main difference between the two is the sound pressure level (*SPL*) and vibration energy ratio. The electret microphone converts only the sound waves into voltage difference, while the contact microphone reacts to the vibration of the metal wall. When somethings hits the box, a branch of a tree, water, rocks, etc., it generates significant vibrational energy compared to the energy that is generated by the corresponding sound pressure waves. In summary, when a mechanical impact happens, the contact microphone's signal gets clipped while the generated acoustic SPL remains low, resulting in small amplitudes in the electret microphone signal. In contrast, when an acoustic wave reaches the device, only a small portion of the energy is converted to vibration resulting in small amplitudes in the piezo microphone signal. However, the electret microphone signal will be pronounced and it may even clip when a high SPL wave reaches the sensor. Based on these observations, knocks can be filtered out efficiently, which is important, because these types of events have impulsive nature and occur frequently in the wild.

The third stage implements the most complex analysis. During the recording phase, preprocessed signal buffers are created, and a number of features are computed from them representing basic relations between several well-defined points of the N-wave pattern.

2.8.2 Shockwave detection

Gunshot detection mainly relies on shockwave classification in the proposed system. During the recording phase, an online algorithm is constantly searching for possible shockwave candidates. It is done by finding consecutive jumps and zero-crosses in the signal and only sections with proper lengths are inserted into a candidate list. Later, only these candidate regions are analyzed, which reduces the processing time.

Figure 2.9 shows a possible shockwave candidate region that has a proper length, mandating further analysis. The procedure starts by finding key points in the pattern, namely the start, the maximum, the middle, the minimum, and the end points of a hypothetical N-wave shape. These points are also illustrated in Figure 2.9.

Based on the well-known shape and symmetry of the shockwave, 10 features are calculated. Let us denote the raw microphone signal by s, a discrete-time signal with the i^{th} sample accessed by $s[i]$. s_0 denotes the same signal after bias level component removal. s_{lin} is the linearized version of s, the result of connecting the key points

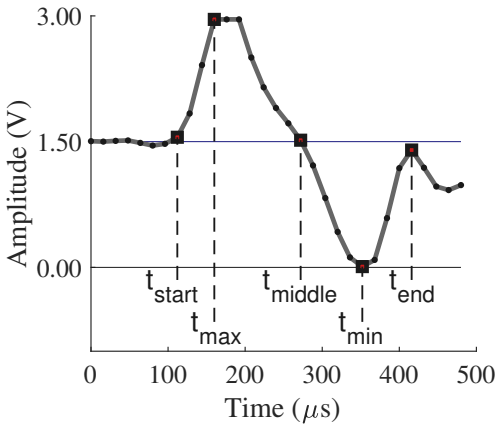


Figure 2.9: *Shockwave candidate with marked key points.*

with straight lines. $\hat{t}_{min}$ denotes the estimated location of the minimum point based only on the first half of the shockwave pattern. Table 2.1 presents the 10 extracted features.

As it will be explained in Section 2.9, a set of shooting range tests were carried out during the development of the algorithm. These recordings and prior knowledge were used to define functions $\phi_i,\ i = 1, .., 10$, which assign [0,1] real-valued numbers to the extracted features $f_i,\ i = 1, .., 10$. The types and parameters of the ϕ_i functions were partially determined by analyzing the distributions of the f_i features in the collected recording set. These functions are similar to the membership functions in the field of Fuzzy theory. However, we do not use rules and the Fuzzy operators, instead a simple aggregation function, σ is defined. The σ function computes the

Table 2.1: *Extracted features* $\mathbf{f_i}, \mathbf{i = 1, .., 10}$ *from the shockwave candidates. These features are based on the known shape and symmetries of the* N*-wave pattern.*

$Maximum\ amplitude = max\{\|s_0[t_{max}]\|, \|s_0[t_{min}]\|\}$
$Symmetry_T = \|(t_{end} - t_{middle}) - (t_{middle} - t_{start})\|$
$Symmetry_{Energy} = \sum_{i=t_{start}}^{t_{end}} s_0[i]$
$Rising\ time_{Start} = t_{max} - t_{start}$
$Linearity_{\{max,middle\}} = \sum_{i=t_{max}}^{t_{middle}} \|s[i] - s_{lin}[i]\|$
$Linearity_{\{middle,min\}} = \sum_{i=t_{middle}}^{t_{min}} \|s[i] - s_{lin}[i]\|$
$Minimum\ point\ error = \|t_{min} - \hat{t}_{min}\|$
$Clipping = \#\ of\ clipped\ points$
$Rising\ time_{End} = t_{end} - t_{min}$
$Amplitude\ symmetry = \|s_0[t_{max}] - s_0[t_{min}]\|$

weighted sum of the feature vector $F = [\phi_i(f_i)]_{i=1,..,10}$ with a weighting vector $W = [w_i]_{i=1,..,10}$, where w_i is the importance of the corresponding feature f_i. The sum of the w_i weights must be equal to 1.0. In that case, the σ function produces a score between 0.0 and 1.0, which reflects the "shockwaveness" of the analyzed signal section. In the current implementation, weights were tuned accordingly to emphasize the importance of the initial section of the N-wave pattern as the end region might be affected by distortions.

2.8.3 Muzzle blast detection and filtering

In contrast to the unique-shaped shockwave pattern, the muzzle blast does not possess any accurately distinguishable signal shape. Different rifles generate different muzzle blast signatures, which depend on many parameters of the barrel, the cartridge and the acoustic environment. With suppressors, the loud, impulsive nature of the sound can be distorted too. Therefore, our system does not rely on the detection of these acoustic events, but may use the information for additional confirmation. The implemented muzzle blast detector analyzes the recorded signal in the time and energy domains. The time domain analysis uses the impulsive nature and loudness of the blast and filters out false patterns by their length. The energy domain detection tries to examine the same, corresponding features. In both domains, only inaccurate recognition is possible, but if both of the methods mark the same section of the signal as a muzzle blast region, then the system interprets this as a possible muzzle blast candidate, which may lead to a final decision with higher confidence.

The structure of the entire recording is also analyzed by a filtering component. For example, if the recording contains a high SPL but periodic, long lasting signal and a less perfect shockwave pattern close to the end of this recorded signal, it is probably a false detection. In contrast, if a less reliable shockwave detection occurs at the very beginning of the recording and a weak muzzle blast detection also occurs later, and the whole signal contains only these two impulsive sections, it is likely that a gunshot event happened. A set of similar intuitive rules have been implemented to strengthen or weaken the detection outcomes.

2.8.4 Final aggregation and examples

The three separate components in the third stage of the gunshot detector (shockwave detection, muzzle blast detection, and filtering) all produce a [0,1] real-valued outcome. These values are combined into one final output, which reflects the probability of a gunshot event. The aggregation emphasizes the importance of the shockwave detection result. However, if the filtering method rejects the recording based on the mentioned intuitive rules, the aggregation is bypassed and the final output becomes

0.0 without the execution of the additional detection algorithms.

The use of soft computing has a benefit of postponing critical decisions to the later stages. In our case, it means that a confidence level can be attached to the gunshot alert and the anti-poaching team can take it into consideration as well as various other factors before a response is initiated. Therefore, our approach is to set a threshold only for alert sending and never make a strict decision about events. The alert sending threshold is not finalized in the current state of development yet; the fine tuning of this part of the system will happen after the completion of the ongoing wildlife tests in Africa (see Section 2.9).

In Figure 2.10,, a set of examples can be seen; the recording of a knock, an animal sound, and a gunshot. It also illustrates the basic structure and behavior of the detector, where the recordings propagate through the stages. In all three cases, the upper signals correspond to the contact microphone, the lower signals to the electret microphone. All of the recordings contain active acoustical events, so Stage 1 lets them through. Stage 2 filters out mechanical impact events, and as it can be observed, the vibrational energy is overwhelming compared to the acoustical energy in the leftmost recording. This ratio between the energies suggest that a physical contact event happened and this recording does not reach the next phase. The two acoustical events are processed by the final stage, where probabilities are assigned to each recording. In the case of the animal sound, the recording does not contain impulsive sections, just periodic pattern, so the filtering method of this stage rejects this signal and interrupts the execution of further detection algorithms. The rightmost

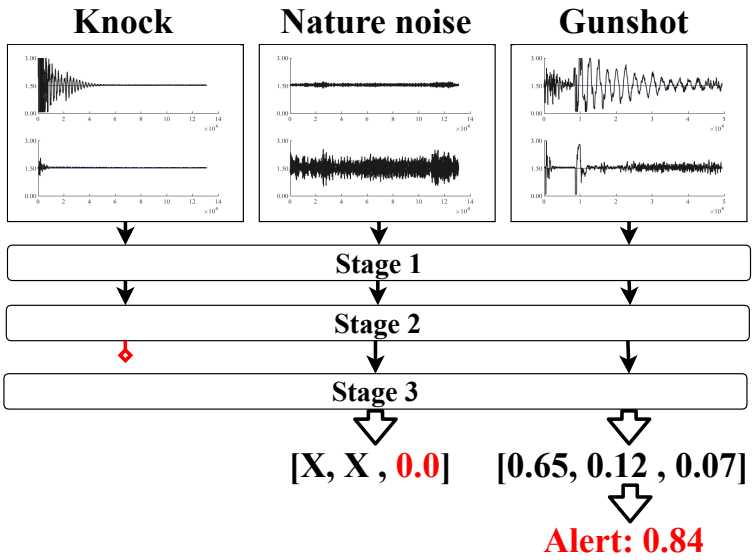


Figure 2.10: *The structure and the behavior of the detector with three example recordings propagating through the stages. In these recordings on the top, the upper signals correspond to the contact microphone, the lower ones to the electret microphone.*

example is a true gunshot, therefore, shockwave and muzzle blast detection happen and the filtering method confirms it by analyzing the whole recording. The output vector is aggregated from these three values and an alert message is sent with a high confidence that a gunshot occurred.

2.9 Evaluation

To help the development of the gunshot detection system, numerous experiments were carried out. Typically, tests were performed on the shooting range to assess and understand the nature of shockwave propagation in different structures. The choice of materials, microphones, amplifications and the fine-tuning of the detection algorithm were all based on these field experiments.

Animal-borne tests were performed too, where we could collect real-world data, sounds, mechanical impacts, accelerometer data, etc. During the first such experiment, the device was worn by a cow for two weeks. Note that this unit did not have the GPS board and it only contained a single battery and a much lighter counterweight. The purpose of this test was to evaluate the mechanical durability of the structure, to check the robustness of basic software components, and to collect real acoustic events through the delay line.

The second real-world experiment became possible with the help of the San Diego Zoo and Safari Park. During this test, elephants wore the device for two weeks. To reduce the load on any one animal, each elephant carried the box for two days at a time. A static node close to the elephant herd was also deployed. The two nodes collected environmental sounds produced by the elephants, neighboring animals and people. The boxes successfully survived the proposed two weeks and collected 7500 events.

The third phase of experiments is currently ongoing in Africa, where prototype sensors are being deployed on wild elephants and they collect wildlife sounds and accelerometer data under real-world conditions. The internal assembly of the unit is presented in Figure 2.11.

To evaluate the detector algorithm with representative data, a set of additional impulsive sound effects have been collected. These events included animal and natural environmental noises such as thunder and rain. The collection was played back in a sound studio and the acoustical events were re-recorded by our device. The playback was carried out by a high-end speaker with a flat frequency curve at high SPL levels. All of the produced events exceeded the 110 dB SPL level, which was verified by a measurement microphone. These pressure levels are rare in the nature but our wake-up mechanism, which is fine-tuned for gunshots, requires the presence of SPL levels above 90 dB to activate. These recordings allow the analysis of the detection algorithm separately, since most of them would be rejected by the earlier stages of

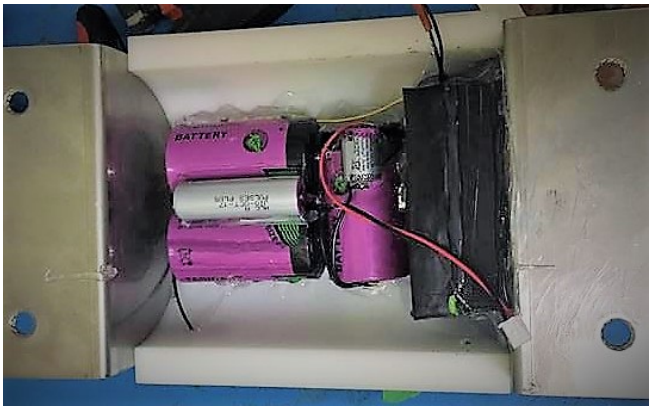


Figure 2.11: *Commercial GPS tracking unit extended with the gunshot detection subsystem before being filled with resin. The protective black box contains the acoustic sensor board and the delay line. One of the batteries (purple cylinders) is reserved for the gunshot detector and the remaining ones for the tracking unit (located underneath). Additional supercapacitors (grey cylinders) are included to help with the current spikes of the GPS and the SD card.*

the detector in their real form.

From the various animal tests and experiments, a dataset has been created. The dataset contains 1000 mechanical impact noises (collected by knocking the box with different materials), 1683 nature sounds from the sound studio experiment, and 7500 events from elephants and from their environment. 72 gunshots were also recorded with the final structure of the device, which are independent from the samples that were used during the development of the gunshot detection algorithm.

2.9.1 Results

Each subgroup of the dataset was analyzed separately. As was explained in the previous section, the detector output represents the probability of gunshot pattern containment in the particular recording. The output values of the detection algorithm were collected into histograms, where the vertical axes are presented on a logarithmic scale, because the comparison of numbers with different magnitudes are required. The results are summarized in Figure 2.12.

In the third (green) row in Figure 2.12, the histogram of the 1000 outputs of the mechanical impacts subgroup is presented. It can be seen that only two inputs received small but greater than zero probabilities. Most of the recorded signals have an impulsive nature and detection based only on the microphone signal would be challenging. However, our method uses the cross-domain filtering in the second stage of the detector, which efficiently filters out these types of events.

The results are similar in the case of nature sounds shown in the second (blue)

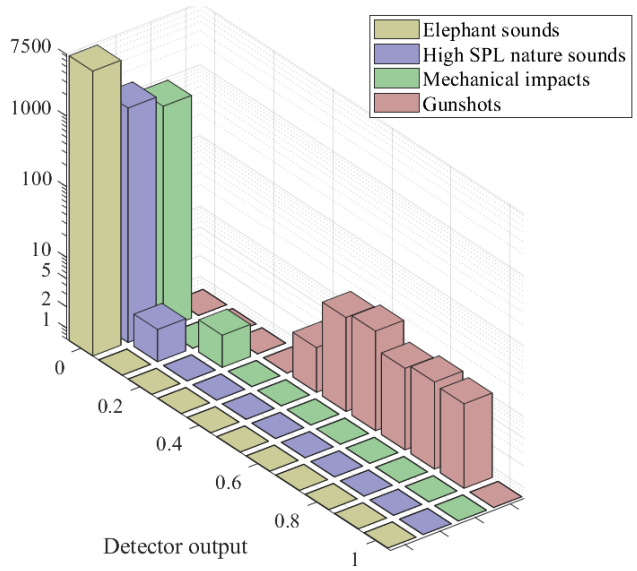


Figure 2.12: *Output histograms of each subgroup of the dataset. A clear separation between the various sounds and gunshots can be observed. Note that all acoustic events analyzed were recorded with the actual sensor.*

row in Figure 2.12. These acoustic events are the most challenging for the classification algorithm, since only the pattern recognition part is responsible for false positive rejection. However, the shockwave and gunshot patterns are unique in nature, which makes the recognition problem feasible with high accuracy. Only two of the 1683 events resulted in (slightly) greater than zero shockwave confidence levels.

The first (yellow) row in Figure 2.12 illustrates the accuracy of the detector, when the input data came from real elephant environment. All of the 7500 events rejected correctly, as no gunshots happened near the elephants. The result is promising, because this group of the dataset contains samples that are probably the closest to real-world wildlife sounds.

We have seen the detector's false positive rejection performance but the main challenge is to remain sensitive to the lower quality gunshots at the same time. The last (red) row in Figure 2.12 presents the result of the detection algorithm on gunshot recordings. As it can be observed, all of the samples have probabilities that are suggesting gunshot activity. On the shooting range, during the tests, we varied the distance between the sensor and the rifles (AK-47 and AR15), and the orientation and distance of the box relative to the bullet trajectory. We also covered the box with mud to simulate real-world conditions. Because of these effects, the recorded shockwaves and muzzle blasts vary in quality.

During the real application of the presented device, false alarms will be expensive and immediately result in loss of trust. Therefore, the alert-sending threshold must be chosen carefully. Since our first wildlife tests are currently ongoing, this function-

ality of the system will be fine-tuned when the data become available. However, from the analyzed dataset, the threshold value can be estimated. The maximal confidence value was 0.22 for non-gunshots, while the lowest gunshot score was 0.44. Therefore, a threshold value between these two numbers would result in 0 false positive alerts, while all of the gunshots would be reported.

2.9.2 Power consumption

The power measurements presented in Section 2.6 were based on a laboratory circuit of the acoustic channel and the microcontroller. The current operational prototype includes power regulators, the accelerometer circuitry and an SD card to save all acoustic and acceleration data to fine-tune the classification algorithms for the next version. The power consumption of this prototype board in different states was also analyzed to estimate the expected lifetime. In sleep mode, it consumes 250 μA, which exceeds the value presented in Figure 2.6 by 150 μA due to the additional components.

In active processing mode, the power consumption is 4.6 mA, which is dominated by the CPU running at a high frequency. Note that the SD card was physically removed during the tests and the corresponding data logging software functions were disabled. The lengths of the active periods are dependent on the level of action required by the detector algorithm. If a false wake-up event happens, Stage 1 immediately sends the device back to sleep mode, and the active mode only lasts for 3 ms. If the wake-up is caused by a mechanical impact or an acoustic event, the length of the signal buffer used is 120 ms. Stage 2 relies only on features calculated online, thus its output is generated instantly and the required duration is therefore 120 ms. If a signal reaches Stage 3, the most complex part of the detector is executed to analyze the possible shockwave and muzzle blast candidates. As the signal buffer is short (8000 samples), and all of the methods have time complexity of $\mathcal{O}(n)$, Stage 3 terminates rapidly. The (over)estimated maximal total time of a gunshot detection is 240 ms.

To estimate the rate of wake-up events, we analyzed our San Diego Zoo dataset. During the two weeks of deployment, the average rate was 20 events/hour including the false wake-up cases. If we assume that all of these events reach Stage 3, and an additional 1000 false wake-up and 100 mechanical impact events happen every hour, which is a safe overestimate, the average power consumption becomes 273 μA. If we double these rates (40, 2000, 200), the current draw only rises by 10% to 298 μA. In the current GPS tracking collar setup, the battery dedicated to the gunshot detector subsystem has a nominal capacity of 19 Ah, which offers a lifespan of 8 years with the estimated event rates. As the proposed lifetime of a tracking collar is only 2 years, smaller batteries could be used, enabling the use on a wide variety of animals.

2.10 Data-driven gunshot detection

The presented classical gunshot detector relies on the N-wave shockwave pattern. Based on this prior knowledge, it was possible to design an accurate classification method. It is well-optimized, the decision-making time interval is comparable to the length of the acoustic event. These properties make the developed method ideal in our system, therefore, it is not highly important in the current phase to examine more sophisticated algorithms that could be used in the gunshot detector. The research is in a data-acquisition phase, our prototype is actively collecting real-world events from wild environments. If there will be events from this dataset that false-trigger the detector, then adjustments will be required. In that case, it is reasonable to proactively scout out relevant directions that could provide more robust and accurate solutions. In this section, today's most popular such methods, neural networks, more precisely, convolutional neural networks are investigated from this perspective. I briefly introduce the concepts used throughout the section and present two approaches, a 1D and a 2D convolutional neural network-based gunshot detector.

2.10.1 Convolutional neural networks

Deep neural networks (DNNs) rule the world of modern pattern recognition and machine learning. These AI algorithms are purely data-driven and learn such data manipulation processes through many consecutive steps, called layers, that lead to proper representations, on which the high-level task can be solved accurately. The nature of the task, the structure of the network, and the layer types make these methods diverse. Their application, mostly in research areas, is widespread and popular software frameworks like Keras [16] or PyTorch [63] make them accessible for researchers coming from various fields.

In this section, Acoustic Event Detection (AED) networks are examined. One of the most widely-used types is called Convolutional Neural Network (CNN), which processes the input through consecutive convolution steps. The advantage of the convolution operation is that the network can be adjusted to match the dimensionality of the input (1D - time series, 2D - images, 3D - volumetric data, etc.). These convolutions are carried out with kernels that are formed and fine-tuned - learned - during the training process. The goal is to evolve such kernels that extract highly-descriptive features to support the succeeding decision-making stages. In our case, it means that the shockwave pattern is not hard-coded into the method but it must be learned from the data.

State-of-the-art solutions develop so quickly that even remaining up-to-date is challenging in these fields. Surveys and comprehensive papers are uncommon, better starting points for interested readers are the summaries of popular data-science

challenges like [14, 58, 66]. The theory of the field is explained in [33] and the practical aspects are covered in [31], and in many other works. More specialized, gunshot detection related approaches can be found in [2, 45, 54]. Because of space limitations, beyond this point, it is assumed that the reader has a basic understanding of the notions and terminology used in deep learning.

2.10.2 Approach

The problem remained the same, so it was to distinguish gunshot recordings from various other events. In this case, the detector needed to learn from the available dataset the regularities between gunshots that are not valid to the other events. In the previous sections, we explained that the N-wave pattern is unique and many false events can be filtered out based on the relationship between the microphone and piezo signals. If a classifier figures out at least these rules, it can achieve a high accuracy.

For the NN experiments, I used the same dataset introduced in Section 2.9. It is imbalanced as it contains only 72 gunshots and thousands of negative samples. To reduce this effect, a subset of the negative examples was selected based on an empirically chosen amplitude threshold. Recordings containing samples above this threshold value were kept. The resulting dataset contained 72 gunshots and 2238 negative samples including real-world elephant noises, various nature noises, and mechanical impact sounds.

The limited number of gunshots opposed a problem during the training and evaluation phases. To make the evaluation more reliable, it is a good practice to employ k-fold cross-validation. The dataset was separated into 6 folds. The models were trained on 4 folds while keeping one fold for validation and one for testing. To calculate a final classification accuracy value, all ”fold-combinations” were tested and the results were averaged. Another problem came-up during the training process as mini-batches were likely to contain only negative samples. To avoid this scenario, batch-creation was controlled manually. The batch-size was fixed to 8 and two gunshot events were picked randomly and inserted into each batch. To further reduce the over-fitting effect of the network on the limited number of gunshots, augmentation techniques were utilized. These included amplification, time-shift, and Gaussian noise addition, which manipulations were carried out on-the-fly during the training. Only the gunshot events were augmented as the negative samples had a proper cardinality and variance. The last adjustment to balance the dataset was the configuration of class weights. In every single batch, there were 6 negative examples and only 2 gunshots, therefore, the class-weight of the latter was set to $3\times$ the class-weight of noises.

The networks were trained by employing the Adam optimizer [48]. The number of epochs was controlled by monitoring the validation accuracy and the training was halted when it had not improved for 10 epochs. The model that achieved the maximal accuracy on the validation set was saved and evaluated on the test dataset.

The target hardware — the gunshot detector — is an embedded device, which can only execute simple, low complexity algorithms in real-time. Therefore, beyond accuracy, other metrics are also important such as computational and memory complexities. The depth of the network, the size and receptive field of the applied convolutional kernels are parameters in CNN structures and are crucial to be chosen correctly to maximize accuracy while keeping the complexities low. However, it is not straightforward how to set these parameters. To support this hyper-parameter tuning phase, a randomized search algorithm was implemented that generated different convolutional neural network architectures with parameters selected randomly from given intervals. This approach is called Neural Architecture Search (NAS) and more sophisticated ways do exist beyond random selection [27]. In our case, the training processes converged quickly as the models were simple and the training dataset was small, and a lot of experiments could be run, therefore, random sampling was a suitable option. The generated models were trained on one fold-combination and then sorted based on accuracy, computational- and memory complexities. After this selection, the best-performing models were evaluated in the 6-fold cross-validation fashion. Additional details will be presented in the forthcoming sections.

The main motivation of this section is to examine several simple methods rather than giving a full picture about the currently available NN solutions, which would be challenging in even a separate work. It is more like a preliminary investigation about what could be achieved with baseline algorithms.

2.10.3 One-dimensional case: Time-domain

In Section 2.8, we presented a gunshot detector that works directly on the raw microphone and piezo signals. These time-domain signals can be fed also into one-dimensional convolutional neural networks (1D-CNN). To do so, the recordings from the dataset were truncated to 2048 from 8192 samples and mapped to a [-1,1] amplitude range. This reduced length is long enough to contain full gunshot events and beneficial as dimensionality reduction speeds up the CNN experiments. The two channels, namely the microphone and piezo signals were encoded into (2048,2)-sized 2-channel tensors. From these pre-processed tensors, the batches were generated manually as was described earlier. One such batch is illustrated in Figure 2.13.

The randomized NAS algorithm was employed to generate diverse 1D-CNN models. These architectures contained 1D convolutional layers followed by 1D maximum pooling layers. After the last maximum pooling layer, the feature maps were concate-

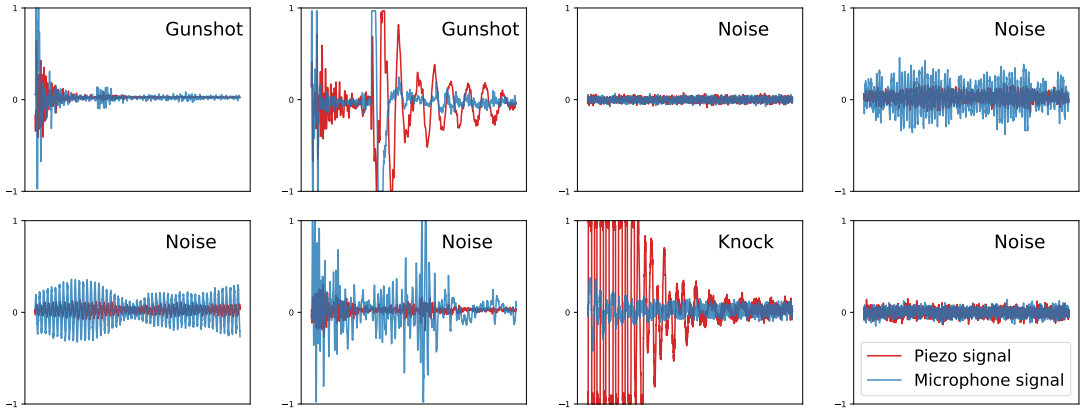


Figure 2.13: *Example batch used during the training of the 1D-CNNs. The microphone and piezo channels are presented on a single plot that makes their comparison easier.*

nated to a single vector, which was forwarded to fully-connected layers. If multiple fully-connected layers were present, Dropout regularization was injected between them. The randomly selected parameters and their optional values are listed in Table 2.2. If, for example, the number of kernels was chosen greater than 1, the dependent parameters became lists, where each list element encoded the parameters of the corresponding convolutional layers.

The NAS algorithm generated 100 valid models. Each model was trained on one combination of the 6-fold partitioning strategy. Four folds were reserved for training, one for validation during training, and one for testing. From the generated model set, the best-performing ones were selected and evaluated on all the fold combinations. Two accuracy metrics were calculated. One of them was the overall accuracy (correctly classified / all samples), which is a good measure of performance. However, in the test fold, there were only 12 gunshots and 373 negative samples. By simply

Table 2.2: *Randomly selected parameters and their optional values in the generated 1D-CNN architectures.*

Parameter	Value
Number of convolutional layers	[1,4) interval
Number of kernels	[2,4,8,16,32] set
Kernel sizes	[3,5,7,9,11,13,17,23,33] set
Dilation rates of kernels	[2,3,4,5,9] set
MaxPooling1D pool sizes	[1,2,3,4,5,7,9] set
Number of fully-connected layers	[1,5) interval
Size of the fully-connected layers	[5,10,15,20,25,35,50,70,100] set
Dropout regularization probabilities	[0.001, 0.2) interval

giving a 'noise' label to all the recordings, a model can achieve 96% overall accuracy. To overcome this problem, the class-wise accuracies were also calculated, and only models with acceptable overall and also with high gunshot-class accuracy — called the recall — were selected.

It was already mentioned that not only the accuracy that matters in the embedded-world. The memory footprint of the model and its computational complexity are also important. These metrics were also determined for all the selected 1D-CNN candidates. Figure 2.14 present the result of the NAS algorithm by illustrating the different metrics of the 10 best-performing models. In each plot, the horizontal axis corresponds to the overall accuracy and the vertical axis to the gunshot class-wise accu-

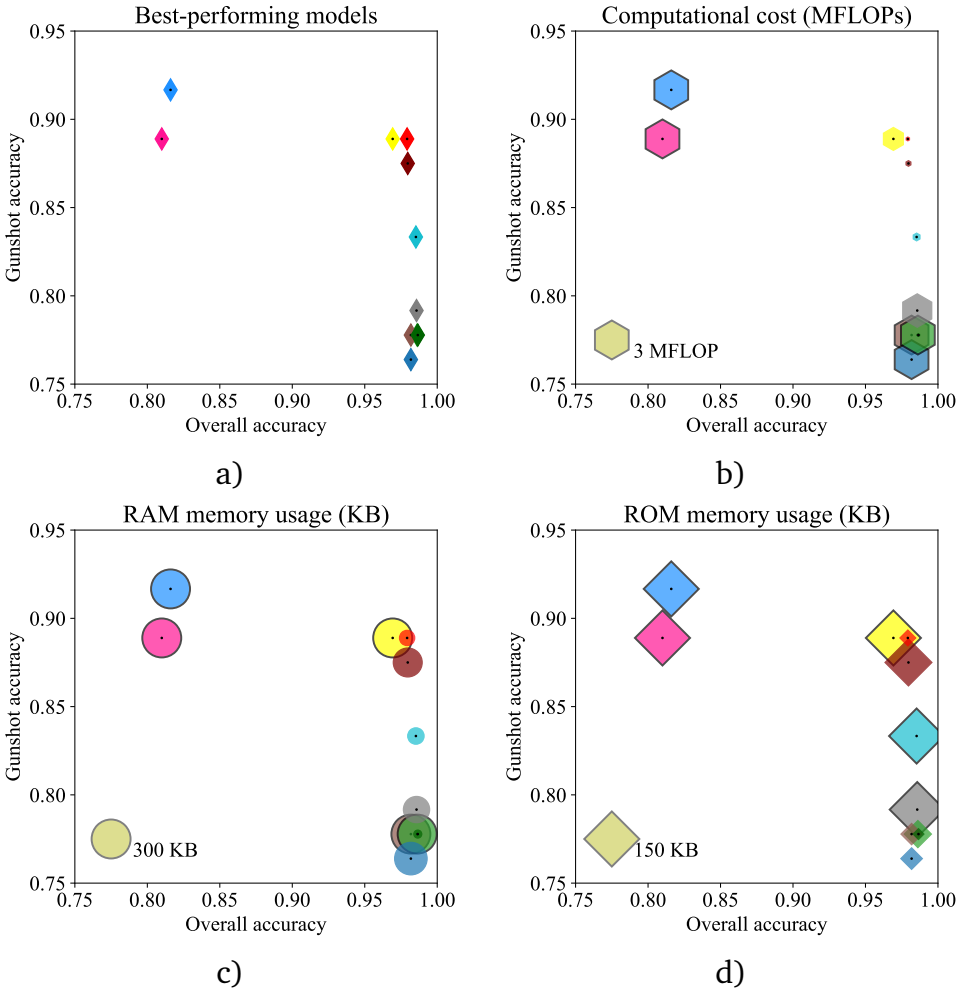


Figure 2.14: *Results of the randomized neural architecture search algorithm. The plots present the different metrics of the best-performing models: a) shows the classification accuracies, b) presents the computational cost of the corresponding structures, c) and d) illustrate the RAM and ROM memory footprints of the models, respectively.*

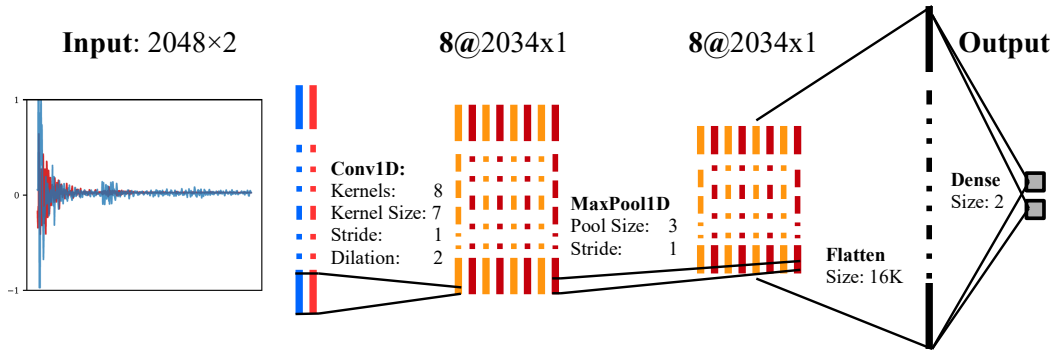


Figure 2.15: *The structure of the best-performing 1D-CNN model.*

racy. Thus, models closer to the top-right corner perform better. Figure 2.14a displays the models' position in the described 2D accuracy space. These models performed similarly but the computations required to produce their results vary, which can be observed in Figure 2.14b, where the diameters of the symbols express the computational cost of the corresponding models. The scale is shown in the bottom-left corner. The solid black edge around a symbol indicates that for visualization purposes the actual values were saturated at the maximal visualized level. Figure 2.14c—d were created with a similar idea but the values expressed by the radii were RAM and ROM memory footprints, respectively.

The best-performing model, the one (with a red symbol) closest to the top-right corner was selected for further analysis. This model achieved 88.8% gunshot classification accuracy and 97.9% overall accuracy. Its architecture is shown in Figure 2.15, It has only one convolutional layer with 8 kernels having sizes of 7 samples with a dilation rate of 2. The receptive field of this kernel is 13 samples ($\approx$200 μs). These extracted features are fed into a maximum pooling layer having a pool-size of 3 samples. The resulting 8 channels are concatenated to a single vector and processed by 2 fully-connected neurons, which produce the final output. The computational complexity of the model is moderate and the RAM and ROM memory footprints are small, thus this solution could be deployed and run in real-time on a microcontroller.

2.10.4 Two-dimensional case: Frequency-domain

Convolutional networks can be easily extended to multiple dimensions. Their primary application was image processing, where the convolutional operators were two-dimensional, thus local features could be extracted from the images. These structures also became popular in AED tasks and achieved state-of-the-art results. In these applications, spectrograms are used to represent the 1D time-domain audio

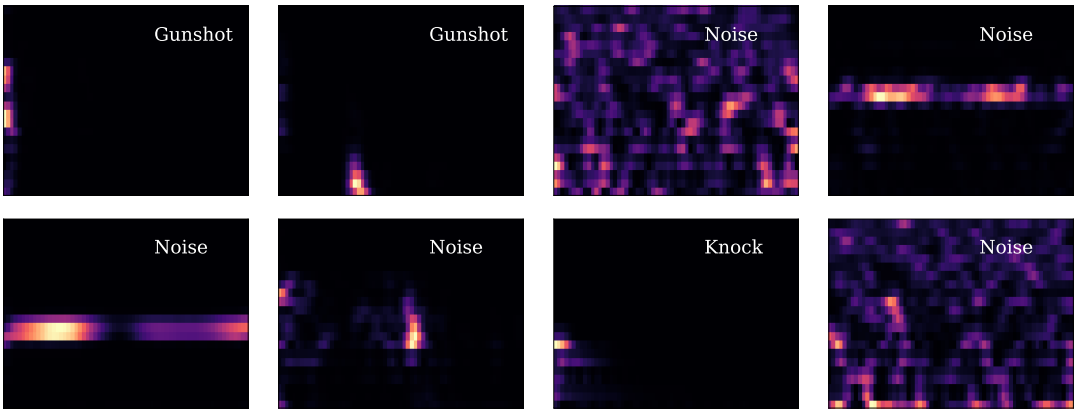


Figure 2.16: *Example batch used during the training of the 2D-CNNs. Only the spectrograms of the microphone signals are presented.*

signals in 2D. The Fast Fourier Transform (FFT) makes the conversion computationally efficient. Mel-scaled spectrograms are also wide-spread, which representation tries to inherit the human ear's non-linear frequency resolution.

The approach and experiments in this section are similar to the previous section, but the trained networks are two-dimensional convolutional neural networks (2D-CNNs). The time-domain signals were transformed to spectrograms by 256-point FFTs with a hop-length of 32 samples. These spectrograms were processed further by mapping them to a Mel filterbank of 22 bands between 100 Hz and 10 kHz. The resulting mel-spectrograms had a shape of 22×65. The microphone and piezo channels were converted to mel-spectrograms separately and then combined to (22,65,2) 2-channel tensors, which served as the input of the models. The batches were generated manually in the previously described way. Figure 2.16 presents a batch, where

Table 2.3: *Randomly selected parameters and their optional values in the generated 2D-CNN architectures. [X]×[Y] denotes the Cartesian product of the sets X and Y.*

Parameter	Value
Number of convolutional layers	[1,4) interval
Number of kernels	[4,8,16,32,64] set
Kernel sizes	[[1,3,5,7]×[1,3,5,7]] set
Dilation rates of kernels	[[1,2,3,4]×[1,2,3,4]] set
MaxPooling2D pool sizes	[[1,2,3,4]×[1,2,3,4]] set
MaxPooling2D stride sizes	[[1,2,3]×[1,2,3]] set
Number of fully-connected layers	[1,5) interval
Size of the fully-connected layers	[5,10,15,20,25,35,50,70,100] set
Dropout regularization probabilities	[0.001, 0.2) interval

only the mel-spectrograms of the microphone signals are shown.

The generation of the 2D-CNNs was slightly more complex than in the 1D case. Each convolutional and pooling kernel had a 2D shape, therefore more parameters should have been chosen randomly. The abstract architecture of the model was similar to the 1D-CNN but the 1D operations were replaced by their 2D counterparts. The full list of the hyper-parameters and their optional values are presented in Table 2.3.

The searching algorithm generated and evaluated 100 valid 2D-CNN models. The best-performing structures were analyzed and illustrated in Figure 2.17. Note that the conversion of the raw signal to mel-spectrogram is not included in the presented values because it is architecture-independent. Globally, these networks performed better than the 1D-CNNs and because of the dimensionality reduction introduced by the spectrograms, their computational and memory footprints tend to be lower.

The best-performing model achieved 98.3% overall accuracy and its recall was

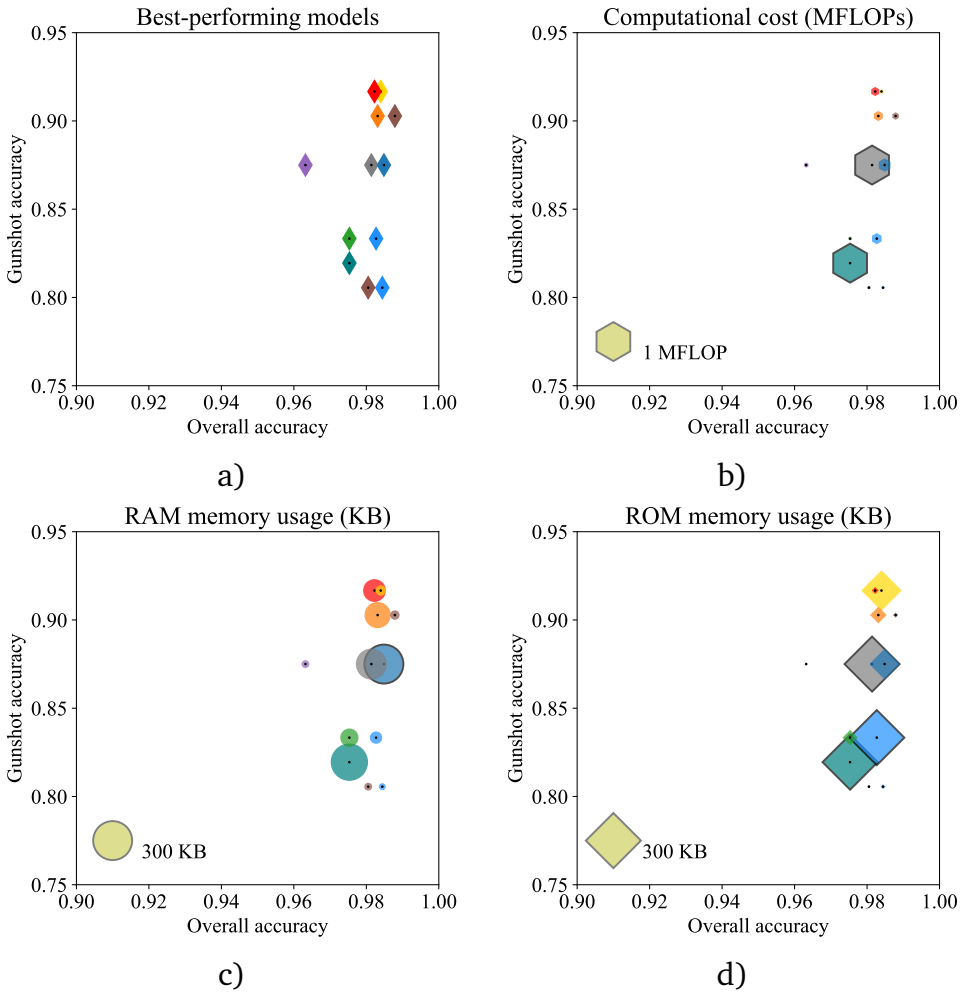


Figure 2.17: *Results of the randomized 2D-CNN architecture search algorithm.*

91.6%, which values are slightly higher than in the 1D case. The selected model's architecture is presented in Figure 2.18. The two-channel input is processed by a 2D convolutional layer. The kernel size is 7×1, which favors frequency resolution that comes from its columnar shape. The produced feature maps are compressed by a maximum pooling layer with a pool size of 3×2 and a stride of 2×1. These maps are reshaped to a single vector and processed by two fully-connected layers with sizes of 15 and 2 respectively. The structure is highly effective in terms of computations and RAM usage, which is favorable in embedded devices. Both of these measures are slightly increased by the extra spectrogram calculation step because the storage of the raw signal is required and FFTs, plus a mapping to the mel-scale must be executed. The ROM footprint is higher, around 200 KB, which is an acceptable trade-off as modern microcontrollers typically have flash memories above 1MB.

Convolutional neural networks were capable of accurate gunshot detection from our data. Even the limited number of gunshot recordings were enough to train the classifiers. However, their performance is weaker and their decision-making process is unclear, which are disadvantages compared to the classical gunshot detection method. From the two investigated directions, the 2D networks performed better, but this might be caused by the extra prior knowledge injected into the system through the utilization of mel-spectrograms. Currently, our viewpoint is to continuously develop and test possible solutions but only to extend and support the classical method. For example, the muzzle blast detection of our classifier could be replaced by a data-driven approach. More data is also needed to increase the models' robustness and to confirm their performance in real-world scenarios.

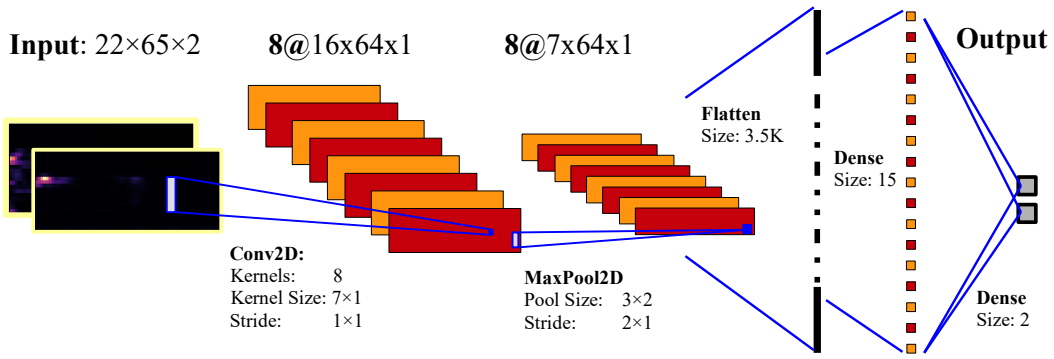


Figure 2.18: *The structure of the best-performing 2D-CNN model.*

2.11 Conclusions and future work

An animal-borne gunshot detection system has been developed to extend currently used GPS tracking collars for elephants. With the fusion of the two systems, gunshot alerts can be raised in real-time coupled with location data. The main challenges were the multi-year lifetime requirement, the preservation of the sound and shockwave quality, and minimizing the false positive rate. With an acoustic delay line structure that utilizes two microphones with different characteristics, the power consumption has been dramatically reduced and the detection accuracy improved significantly. Real-world tests were carried out, including with elephants in a safari park. The collected dataset contained various environmental sounds, mechanical impacts, and real gunshots. The evaluation of the detection algorithm on this dataset showed promising results. Data-driven methods were briefly mentioned and baseline 1D and 2D convolutional neural networks were trained to investigate this direction but their performance remained lower compared to the classical pattern recognition algorithm's accuracy.

The first prototype sensor with on-board storage has been integrated into a GPS tracking collar, and is currently under real wildlife testing in Africa. The collected data from this test will be used for the fine-tuning of our gunshot detection algorithm. The longer term goal of the project is to release all hardware and software in open source form so that gunshot detection capability can be freely integrated into any tracking collars.

2.12 Contributions

The author of this PhD thesis is responsible for the following main contributions

I/1. I proposed a novel acoustic delay line wake-up mechanism, implemented an experimental hardware, and showed that it can improve the power-consumption efficiency of audio event detectors.

I/2. I designed and implemented the hardware and software of an embedded gunshot detector module that utilizes the proposed wake-up mechanism and can be integrated into widely-used GPS tracking collars.

I/3. I developed a novel gunshot detector algorithm that employs the two-domain audio information used for the proposed wake-up mechanism, and evaluated its accuracy and efficiency through real-world experiments.

I/4. I developed a randomized architecture-search algorithm that generated, trained, and compared 1D and 2D convolutional neural networks that utilize the two-domain audio information for gunshot detection.

Chapter 3

Reverse Mode Speakers

In this chapter, the utilization of loudspeakers in acoustic event detection scenarios is investigated. It is well-known that these sound-radiating devices are capable of capturing sounds. which behavior is referred to as reverse mode. This functionality is examined from theoretical and practical viewpoints. First, the electro-mechanical equivalent circuits are formed to analyze the properties of the reverse mode. Then, simulation-based results are presented that explore the event detection capabilities and its limitations. This examination suggests that speakers are able to record sounds with acceptable quality, therefore, a proof-of-concept embedded device is introduced that performs event detection in the inactive periods of speakers. Data-driven methods and possible detection scenarios related to this application are also investigated. A more challenging setup referred to as active reverse mode is also explored and preliminary results are included.

3.1 Introduction

The loudspeaker is an electroacoustic transducer that converts an electrical signal into sound. The most widely used type is the dynamic loudspeaker, which produces sound by forcing a coil with an attached diaphragm to move rapidly back and forth.

The sound generation starts with the electrical signal to mechanical movement conversion. An electrical audio signal is applied to a suspended moving-coil placed in a gap surrounded with strong permanent magnets. This electrical current induces varying magnetic field according to Faraday's law of inductance and a varying mechanical force is being formed by the interaction of the two magnetic fields. The generated force moves the coil and the attached lightweight cone, which oscillation produces acoustic pressure waves. This traditional mechanism of speakers is referred to as the *direct mode*.

However, it is well-known that speakers can record sound as well, and this 'microphone' mode can be found in the literature referred to as *reverse mode*. This reversed

behavior is similar to the working principle of dynamic microphones. The incoming sound waves exert force on the surface of the diaphragm, which starts vibrating, inferring the oscillation of the coil in the magnetic field. As magnetic field fluctuations occur through the coil, electromotive force is being generated, i.e. a voltage difference builds up between the coil's two terminals. This varying voltage represents the incoming sound in the electrical domain.

In this chapter, the reverse mode is analyzed through theoretical modeling and proof-of-concept experiments. The idea is that reverse mode speakers could be utilized in many places and the hardware extension is minimal as it only requires simple modifications. The original, direct mode function is provided but extra functionality becomes possible. For example, they could be used in security applications, where suspicious acoustic event detection is required.

Only passive loudspeakers — the ones without active built-in amplifiers — are examined in this work, because the voltage generated in reverse mode cannot reach the driving cable in active speakers. However, any system proposed here could be integrated into active speakers as well, which would measure the reverse mode signal between the coil and the integrated amplifier.

It will be presented that the audio recording with loudspeakers is sub-optimal compared to the microphone-based solutions. However, if a deployed system already contains multiple, spatially separated but connected passive speakers that are driven by the same source, the whole area could be covered by monitoring the driving cable with a single device. The advantage is the ease of deployment and if the speakers are mostly inactive, event detection is achievable during these periods. This setup is commonly used, for example, in schools, stations, hospitals, etc., where the speakers only occasionally broadcast announcements and remain inactive between them.

A more challenging perspective of the reverse mode utilization is when the speakers are actively driven by a driving source. The task is to detect external acoustic events on the driving line of a loudspeaker while an active driving signal is also present on this line. This detector device, therefore, has access only to the superposition of the dominating driving signal and the weak reverse mode signal. Preliminary investigations about this setup are also included in this chapter.

Structure of the chapter: Section 3.2 summarizes the related works and briefly presents previous results. The analysis of the reverse mode extended with experiments and measurements is covered in Section 3.3. Based on these investigations, the possible applications of speakers in acoustic event detection tasks are included in Section 3.4, where a proof-of-concept device implementing clap detection is also introduced. In Section 3.5, the challenging active reverse mode problem is analyzed. Result and final thoughts are summarized in Section 3.6.

3.2 Related works

Before the research activity of the author, reverse mode speakers had mainly been considered in spying applications as they offer a feasible way to bypass security systems. Thus, there are no previous works related closely to the 'positive' utilization of speakers, which is the main contribution of the current chapter. In the following, several approaches are listed and summarized that employ reverse mode speakers in various applications. More general works related to specific topics are detailed in the corresponding sections with references included.

In recent years, an interesting work related to the reverse mode behavior was published in [36]. They developed a malware, named *Speake(a)r* that allowed using a headphone as a microphone after port retasking, which could be easily realized in many modern PCs. The software altered the functionality of output ports and turned them into input ports. Thus, a headphone connected to the output line became an input device. Experiments were carried out, where they recorded normal conversations from close ranges with acceptable quality. Therefore, they showed that eavesdropping and cyber-attacks are feasible.

Another work [50] presented and analyzed threats that try to leak information from separated networks to internet accessed networks through covert channels. They utilized microphones and loudspeakers to transmit information through the air with acoustic signals above the human audible range between a speaker and a recording device (air-gapped system). Two of the three investigated cases involved speakers in reverse mode as the recording device. They achieved 8 bit/s transmission speed between loudspeakers, which showed that air-gapped systems are not completely secure and an attacker might be able to access critical information from a separate network through speakers.

The authors of [7, 8] introduced a novel framework for tap detection in smartphones, called *TouchSpeaker*. The system exploits the built-in loudspeakers as primary *sensors*. The proposed method combines the signals recorded by the speakers in reverse mode with additionally available sensor readings to achieve state-of-the-art tap detection and classification performance. They could identify nine different types of taps. It was reported that loudspeakers have advantages like low power consumption and high sensitivity for finger taps in noisy acoustic environments. Their results suggest a new opportunity to employ the speakers in smartphones, which direction is also mentioned in Section 3.4, where the localization of the phones, or their users, would become possible with the utilization of the built-in speakers.

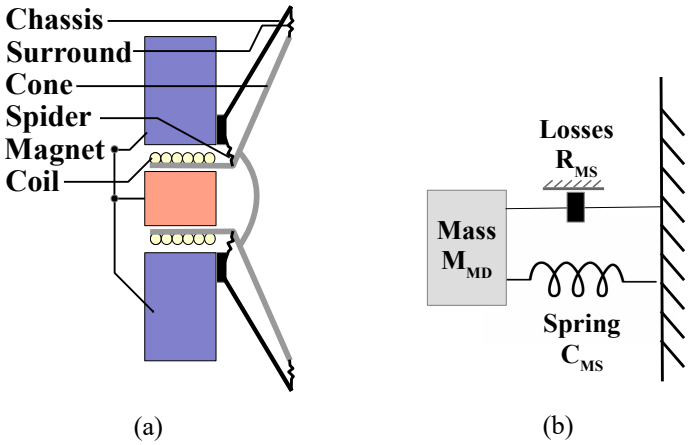


Figure 3.1: *Structure of a moving-coil speaker (a), and its mechanical approximation with a mass-spring-damper system (b).*

3.3 Theoretical modeling

A moving-coil speaker contains a suspended coil placed in a gap between permanent magnets. When an alternating current flows through the wire, magnetic field is being induced, which interacts with the permanent magnetic field. These forming forces move the coil and the attached diaphragm - the cone - back and forth. That rapid movement of the cone generates pressure waves in the air. The cross-section of a loudspeaker can be observed in Figure 3.1a. The explained structure can be approximated by a mechanical mass-spring-damper system presented in Figure 3.1b, where M_{MD} is the mass of the moving parts, C_{MS} is the compliance of the suspension and R_{MS} is responsible for the additional losses.

3.3.1 Equivalent circuit of the direct mode

An amplifier drives the loudspeaker in the electrical domain that induces mechanical force, which moves the cone, producing waves in the acoustic domain. This complex system can be modeled with an equivalent circuit presented in Figure 3.2a by using electrical impedance, mechanical mobility, and acoustic impedance [6, 42].

In Figure 3.2a, a voltage generator with neglected output impedance models the driving source of the loudspeaker, typically it is an amplifier. The coil resistance, R_e is represented by a resistor while its inductance is negligible at the relevant sound frequencies.

It is known that the force on the coil is given by the product of the flux density in the gap, B (T), the length of the coil wire l (m) and the current (A). The constant that connects the electrical and mechanical domains is the Bl product. This parameter is

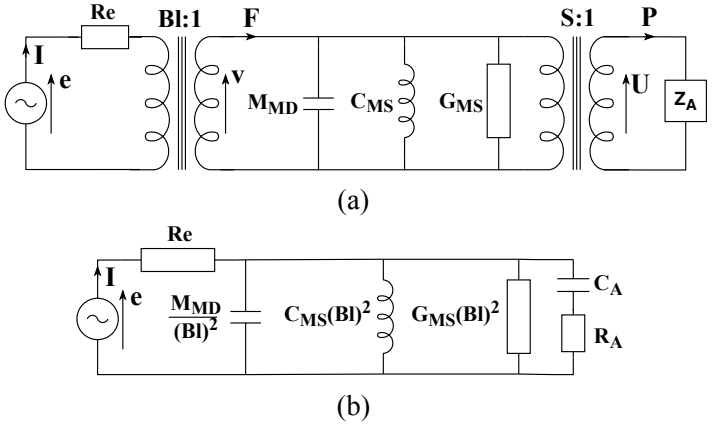


Figure 3.2: *Equivalent circuits of a moving-coil speaker. In (a), virtual transformers are used to model the mechanical force generation and acoustical energy radiation. In (b), these transformers are eliminated, and the loads are transformed into the electrical domain..*

usually given in the loudspeakers' datasheets. This mechanical force generation and impedance transformation can be modeled by a virtual transformer with turn ratio of *Bl*:1. The impedance connected to the secondary of this transformer is reflected back to the primary side by this Bl ratio.

In the mechanical domain, the mechanical mobility analogy is used to model the mass M_{MD}, the compliance C_{MS}, and the losses G_{MS} of the mass-spring-damper system.

The air load has two main parts, the reactive (C_A) and the resistive (R_A) parts. The radiated sound energy is proportional to the square of the diaphragm area S. This load is modeled by applying the mechanical velocity on the primary of another virtual transformer with a turn ratio of S, as shown in Figure 3.2a. On the secondary side, the voltage is proportional to the volume velocity U and the current is proportional to the sound pressure P.

The first transformer can be eliminated by bringing the mechanical load to the primary side with impedance inversion and conversion. The second transformer can be eliminated in the same way, but it does not invert the impedance. After these eliminations, the equivalent circuit will be transformed into the electrical domain, where all the components are replaced by equivalent electrical ones. This simplified model can be observed in Figure 3.2b, where the component values can be calculated from the Thiele-Small parameters [80]. These electromechanical parameters define the low-frequency behavior of a speaker unit and usually are used for enclosure design. They are measured and published by the speaker manufacturers. For example, such parameters are the resonance frequency, suspension equivalent air volume, DC resistance, moving mass, effective diaphragm diameter, etc. The complete list of the

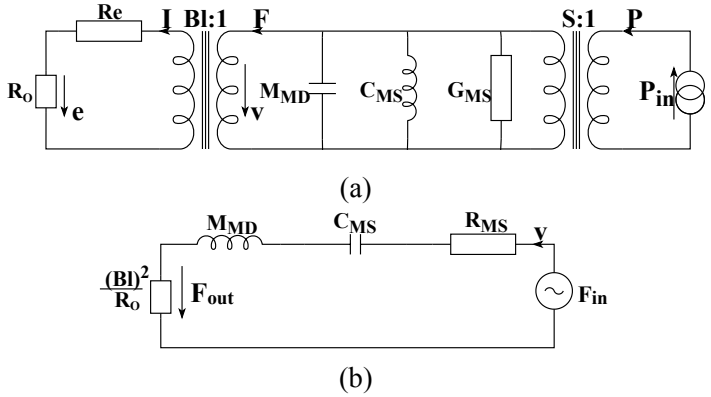


Figure 3.3: *Equivalent circuits of the reverse mode. In (a), virtual transformers are used to represent the force and electrical signal generation steps. The driving source is moved to the acoustic domain. The electrical output signal can be measured on R_O. In (b), the simplified mechanical equivalent circuit is presented after the elimination of the virtual transformers. The elements are transformed into the mechanical domain by using mechanical impedance.*

Thiele-Small parameters and the relations with the values required by the model in Figure 3.2b can be found in [80].

Interested readers may find further details about the loudspeaker models in the literature [6, 9, 29, 42, 80].

3.3.2 Equivalent circuit of the reverse mode

In reverse mode, the mechanical properties of the speaker remain the same, but the excitation forces are originated from the acoustic pressure waves. When an external pressure signal is applied on the surface of the diaphragm, it starts vibrating, and the attached coil oscillates in the magnetic field. According to Faraday's law of inductance, voltage is generated in the coil. This can be modeled in the same way as it was presented in Figure 3.2a, but the driving source is moved into the acoustic domain. As the measurement of the reverse mode voltage is required, the electrical voltage generator is replaced by a resistor, which represents the input impedance R_O of a voltage meter or operational amplifier. The resulting circuit can be examined in Figure 3.3a. The acoustic domain resistance and reactance are eliminated as the driving pressure directly acts on the surface of the diaphragm. The mechanical domain model remains the same but the excitation comes from the other direction.

To simulate a speaker's response to a given acoustic signal, the model is converted into the mechanical domain by eliminating the two virtual transformers. The resistance of the coil, R_e, is negligible compared to the input impedance of the voltage meter ($R_e << R_O$), therefore it is omitted.

The acoustical driving source is a volume velocity signal. The relation between this signal and the mechanical velocity of the diaphragm is determined by the surface area of the cone, the transmission coefficient of the cone material and the shape of the cone. The exact description is complex but an approximation can be made as the U volume velocity can be brought to the other side of the virtual transformer by using the turn ratio.

After the eliminations of the virtual transformers and the forming of the mechanical impedance type analogy circuit, the resulting mechanical equivalent circuit of the reverse mode can be observed in Figure 3.3b. From the simplified equivalent circuit, the reverse mode transfer function of a loudspeaker can be derived as:

$$H(\boldsymbol{s}) = \frac{R_O C_{MS} \cdot \boldsymbol{s}}{M_{MD} C_{MS} \cdot \boldsymbol{s}^2 + (R_O + R_{MS}) C_{MS} \cdot \boldsymbol{s} + 1}.$$

The component values are calculated from the Thiele-Small parameters. The derived transfer function has a band-pass filter nature, which is similar to the measured electrical impedance curves usually included in the speakers' datasheets. This similarity is reasonable as the impedance seen by the electrical driving unit is also dominated by the mechanical properties of the speakers. One example of a simulated transfer function is presented in Figure 3.4. The frequency, where the transfer function has its maximum is called the speaker's resonance frequency, f_s. This is the frequency where the cone is vibrating with the maximal amplitude and velocity, and the generated counter-electromotive force is also maximal, which causes the effective electrical impedance of the speaker to be at its maximum at f_s, At this frequency, the mechanical system has its minimal mechanical (motional) impedance, therefore, an external force with the frequency of f_s leads to maximal perturbations of the di-

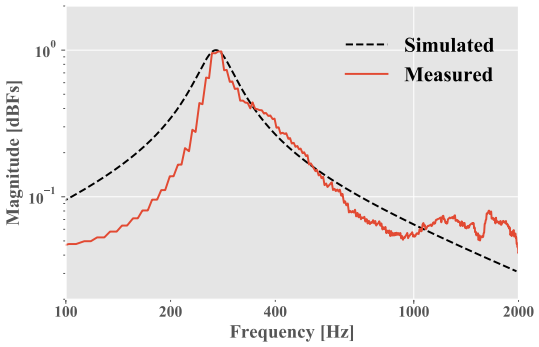


Figure 3.4: *The real and simulated reverse mode transfer functions of a 2" loudspeaker with a resonance frequency of f_s =270 Hz. The maximal values of the curves are located at the mechanical resonance frequency of the investigated speaker. It can be observed that the simulated curve has similar characteristics as the measured curve.*

aphragm, and to maximal generated voltage. In summary, a reverse mode speaker has maximal sensitivity at fs, which is also observable in Figure 3.4.

3.3.3 Experimental results

In this section, measurement results are presented to validate the theoretical results and to further analyze the reverse mode behavior. It also explains and illustrates the steps of reverse mode simulations, which convert traditional recordings into forms as they would have been recorded by loudspeakers.

Transfer function: In Section 3.3.1, the transfer function of the reverse mode was derived. This function is plotted in Figure 3.4, where the frequency curve corresponds to a full-range speaker with 2" cone diameter and a resonance frequency of $f_s = 270$ Hz. Figure 3.4 illustrates also the real, measured version of the curve. The reverse mode transfer function measurement process of the speaker involved white noise as the input signal produced by another direct mode speaker. Before the reverse mode measurement, the radiating speaker was calibrated by a measurement microphone and equalized by software to produce a flat spectrum. Then, a 10-minute-long recording was carried out by the reverse mode speaker. The recorded signal was separated into non-overlapping sections. For each section, the 2048-point FFT spectrum was calculated and these spectra were averaged to obtain the final transfer function curve. The simulated and measured curves can be compared in Figure 3.4. Note that below 200 Hz, the radiating speaker could not reproduce the frequency components accurately, which contributed to the increased error in this range. Furthermore, it is reasonable that with such a simple lumped-element model the exact simulation of the reverse mode behavior is impossible but the characteristics of the simulated curve are similar to the measured curve, thus, simulations based on this transfer function produce comparable results.

Sensitivity: Microphone sensitivity is typically measured with a 1 kHz sine wave at a 94 dB_{SPL} sound pressure level (SPL), which corresponds to 1 Pascal (Pa) pressure. The amplitude of the output signal from the microphone with this input stimulus is a measure of its sensitivity. This parameter is often expressed in the form of mV/Pa, or on a logarithmic scale dBV, dB respect to 1V.

Traditional microphones have more or less flat frequency response curves on wide frequency ranges. Therefore, the measurement at 1 kHz is a good estimate of the sensitivity for the whole covered frequency range. Contrary, the transfer function of a reverse mode loudspeaker has a band-pass nature with a peak at its resonance frequency f_s. The sensitivity at this frequency is much higher than at the other points of the frequency range. Therefore, the sensitivity of reverse mode loudspeakers should be measured at two points: at the resonance frequency, and at 1 kHz to remain comparable to microphones.

During the experiments, a 2", 8 Ω full-range loudspeaker's reverse mode sensitiv-

ity was measured at its resonance frequency f_s =270 Hz, and at 1 kHz. In both cases, the measurements were carried out at 94 dB_{SPL}, validated by a measurement microphone. At f_s, the sensitivity reached 3.53 mV/Pa (-49 dB re 1 V/Pa), and at 1 kHz, it reduced to 0.23 mV/Pa (-72 dB re 1 V/Pa). The sensitivity value around the f_s frequency is comparable to traditional dynamic microphones' sensitivity, which typically varies in the range of 1—6 mV/Pa. At 1 kHz, the sensitivity dropped significantly and this reduction is responsible for the information loss at the higher frequencies.

This frequency-dependent sensitivity characterizes all reverse mode speakers, but they are all capable of recording environmental sound events, or at least, the events' frequency components around f_s. An interesting (or maybe suspicious?) correlation can be highlighted here: full-range speakers (most commonly used ones) usually have a resonance frequency in the range of fundamental speech frequencies (80-250 Hz). If the speakers are located close enough to conversations, they can record the main frequency components of speech similarly to microphones.

3.3.4 Reverse mode simulations

No audio datasets are available, where the events were recorded by reverse mode loudspeakers. Therefore, one needs to simulate reverse mode responses to acoustic events to be able to test and measure reverse mode performance in event detection and classification tasks. These simulations take traditionally recorded audio files as inputs, and with the help of the reverse mode transfer functions, the output, i.e. the response of the speaker to the acoustic event can be estimated.

The simplest simulation technique is to calculate the reverse mode responses as the output of a Linear Time-Invariant (LTI) system characterized by the reverse mode transfer function of a speaker. The input signal of the system is the microphone-recorded version of an event and the output is produced by simulating the system's behaviour to this given input excitation. The resulting signal carries similar characteristics as it would have been recorded directly by the simulated speaker.

In Figure 3.5a the spectrogram of a microphone-recorded audio event is shown. In Figure 3.5b, the same event, recorded by a 2" full-range speaker with $f_s = 270$ Hz is presented, where the loss of information at higher frequencies and the increased noise can be observed. The spectrogram of the LTI system based simulation response to Figure 3.5a is visible in Figure 3.5c. The system is defined by the reverse mode transfer function of the loudspeaker. White Gaussian-noise could be added to the output to simulate the low signal-to-noise ratio, which step is omitted here.

Modern sound classification algorithms rely on deep-learning models and methods. These techniques usually treat sound as an image by converting it to a spectrogram (or Mel-scaled spectrogram). Convolutional networks [52, 55] form a group of neural networks that work on multidimensional input data and process it through

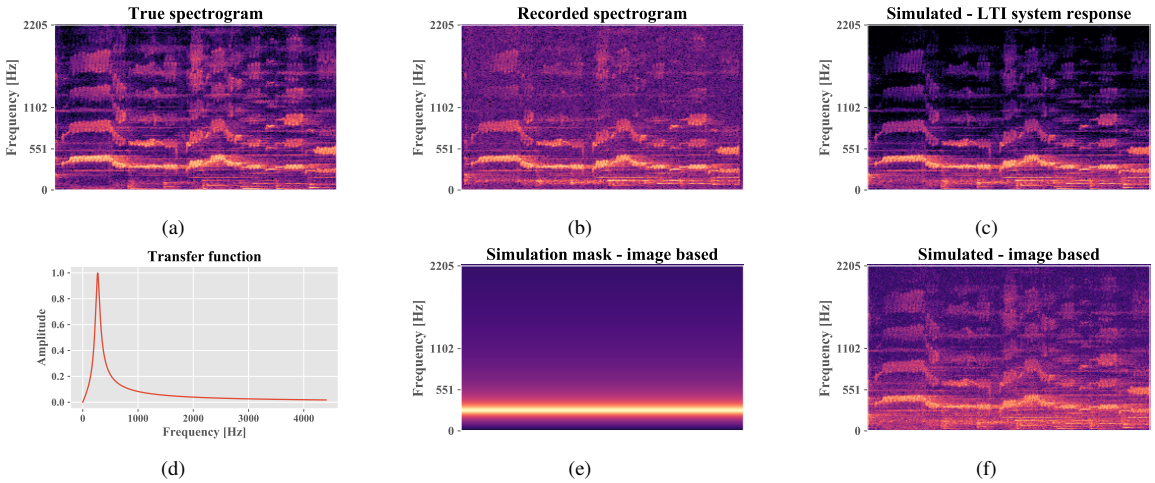


Figure 3.5: *Illustration of the steps of reverse mode simulations. The input spectrogram is presented in (a). In (b), the same acoustic event's spectrogram is shown, but it was recorded by a reverse mode speaker. The goal of the simulation is to produce a spectrogram similar to (b) from (a). The output spectrogram of an LTI system response based simulation can be observed in (c). The steps of a more efficient, image-based simulation method are presented in the bottom row (d)-(f). Based on the reverse mode transfer function (d) of the simulated speaker, a simulation mask is generated (e). By using this mask, the simulation output (f) can be calculated with a simple point-wise multiplication of (a) and (e). To produce a more realistic spectrogram, noise was added to (f), thus the final result is similar to (b).*

consecutive convolutional operations with kernels that were shaped and formed during the supervised training process. The structure learns and extracts high-level features from the data, thus classification based on these features becomes feasible. In the field of sound classification, convolutional networks have played a dominant role in recent years [35, 39, 65].

If one assumes that the classifier will work on spectrograms, a more 'clever' simulation with enhanced efficiency can be achieved. It takes the magnitude curve of the transfer function, similar to the one presented in Figure 3.5d and replicates it as columns to produce a reverse mode mask image shown in Figure 3.5e. With this mask, the simulation complexity can be reduced to a simple point-wise matrix multiplication, where the original spectrogram and the reverse mode mask should be multiplied. One such example spectrogram image is illustrated in Figure 3.5f, where Figure 3.5a and Figure 3.5e were multiplied. Additional noise was applied to simulate the signal-to-noise ratio drop caused by the low reverse mode sensitivity. The visual similarity between Figure 3.5b, Figure 3.5c , and Figure 3.5f can be observed, which indicates the correctness of the reverse mode simulations.

3.4 Utilization

As was mentioned in Section 3.2, reverse mode speakers were mainly examined in security threat scenarios before. However, in this chapter, the 'positive' utilization possibilities are highlighted, which include simulation-based acoustic event detection results, several application opportunities, and a proof-of-concept device called Smart Speaker is explained through the realization of a reverse mode speaker based clap detector.

3.4.1 Urban sound classification: Simulation results

A potential application of reverse mode speakers is acoustic event detection, e.g. suspicious event detection. To investigate what can be expected in such situations theoretically before further analysis, simulation-based experiments were carried out. These examinations simulated reverse mode speaker responses to microphone-recorded events and ran classification algorithms on them to show the effects of the information loss introduced by the low and non-linear reverse mode sensitivity. As no speaker-recorded datasets are available for testing, a dataset recorded by microphones was used and the responses of speakers to these inputs were simulated by using their reverse mode transfer functions $H(s)$.

The original sound pressure levels of the input recordings, the amplification factors, and microphone types, etc. were unknown, thus no exact simulation was possible. To simulate the effect of different input intensities, noises with different power levels were added to the response signals. These levels were: no noise, -40 dB, -30 dB, and -10 dB. High noise power lowers the SNR, thus lower input sound intensities were simulated. White Gaussian noise was applied with zero mean and the variance was set according to the required noise power level. The simulations were performed using the LTI system based method explained in Section 3.3.4.

A dataset called UrbanSound8k was used [75], which contains 8732 labeled sound excerpts ($\leq 4\,s$) of urban sounds from 10 classes: *air conditioner, car horn, children playing, dog bark, drilling, enginge idling, gunshot, jackhammer, siren, street music*. The audio files had various lengths, sample rates, bit-depths and number of channels. To unify these parameters, all of the recordings were converted to a single-channel, 16-bit format with a sample rate of 22.05 kHz. First, each signal went through the reverse mode simulation and then they were split into overlapping segments with lengths of 0.95 s. The same preprocessing was carried out in [13, 65]. Each segment of a signal received the original signal's label. These labeled, transformed segments were used during the classification phase.

As was described in Chapter 2, modern classification algorithms rely on deep learning and neural networks. The best results were achieved by using these methods

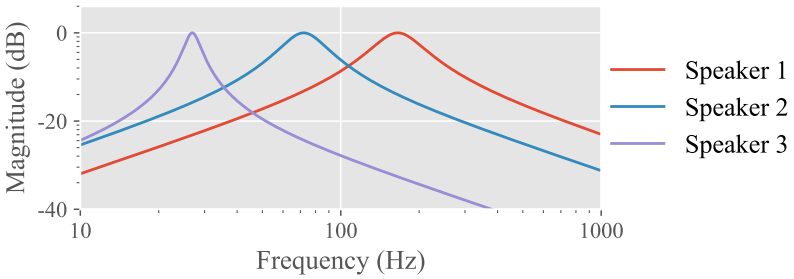


Figure 3.6: *Reverse mode transfer functions of three different loudspeakers. The band-pass filter nature of the curves can be observed. Amplifications were applied to set the magnitudes of the transfer functions to unity at the resonance frequencies.*

in the field of sound classification, too [35, 55]. During the experiments, a convolutional neural network was trained with a structure similar to the one published in [74]. The input format was changed; log-scaled mel-spectrograms were used (as it was presented in [13, 65]).

Three different speakers were tested with 5 cm , 10 cm, and 20 cm cone diameters. These speakers are referred to as Speaker 1 [22], Speaker 2 [23], and Speaker 3 [24], respectively. Based on their published Thiele-Small parameters, their reverse mode transfer functions were calculated and the resulting functions are presented in Figure 3.6. The whole audio dataset was transformed four times per speaker with the four noise power levels described earlier. That resulted in 12 transformed datasets.

The classifiers were trained on the 12 transformed datasets separately until convergence (Early-Stopping was employed based on the validation loss). The dataset was organized into 10 folds. 8 folds were used for training, 1 for validation and 1 for testing. To simplify the training method, instead of the required 10-fold cross-validation scheme, only one training was carried out on each dataset with the same folds selected for training, validation and testing ([1.,8.], 9., and 10., respectively). Thus, the results were comparable and the examination did not require the $10\times$ repetition of the training processes. The baseline accuracy was determined similarly on the original data, in which case the reverse mode simulation step was skipped. The optimal learning rate parameter of the training procedure was determined on the baseline system first, and then the same value was set during all the other trainings. Further optimization could be made in each separate case, which might increase the accuracy, however, these fine-tuning steps are beyond the interest of the current examination.

As loudspeakers are sub-optimal microphones, therefore, it is foreseeable that their reduced sensitivity and frequency selectivity affect the classification accuracy negatively. Table 3.1 summarizes the resulting accuracies in all the tested cases. The accuracy on the original dataset was 70%, which seems close to the state-of-

Table 3.1: *Accuracies of the trained classifiers. In the columns the noise levels, in the rows the baseline setup, and the different speakers are presented.*

	No noise added	**Noise: -40dB**	**Noise: -30dB**	**Noise: -10dB**
Original	70%	-	-	-
Speaker 1	61%	60%	60%	50%
Speaker 2	56%	54%	55%	42%
Speaker 3	56%	58%	53%	32%

the-art [13], however, no exact comparison can be made, because the 10-fold cross validation scheme was violated in this work. Still, the results are comparable within the table, since all the trainings and testings were carried out on the same, but differently transformed datasets.

It can be observed in Table 3.1 that the speaker-based accuracies are at least with 9% worse than the baseline accuracy. This is mainly originated from the nature of the reverse mode frequency response. It is noticeable that the higher the diameter and lower the resonance frequency, the lower the classification accuracy becomes. Speaker 1 achieved the best performance, which can be explained by its highest resonance frequency, thus relevant frequencies are less attenuated. The noise level also had an impact, however, it was not negative in reasonable ranges (-40dB, -30dB). This is not surprising; adding noise to signals is a commonly used data augmentation method [74].

To better reveal the source of accuracy drops caused by the reverse mode speakers, in Figure 3.7 the difference between two confusion matrices are presented: *(baseline matrix - Speaker 1, no noise matrix)*. Interestingly, the loud, impulsive events like *gunshots*, *dog bark*, and *car horn* were well-classified with the reverse mode speaker based classifier as well. Most of the error came from the less intense events with periodic nature like *drilling*, *air conditioner*, and *engine idling*. With the high frequencies attenuated, these periodical events produced similar spectrograms.

From the classification results, it can be concluded that reverse mode speakers could be used for event detection. However, the type and nature of these events are limited. For example, reliable speech recognition could hardly be achieved because of the low sound pressure levels. At the same time, loud, impulsive events like gunshots, explosions, screaming, etc. could be detected with sufficient accuracy. These observations can be explained based on the analysis results presented in Section 3.3. The reverse mode sensitivity is low outside the f_s resonance frequency zone, therefore the information content of loud events is more likely to be preserved. Impulsive events have wide coverage in the spectrum, thus at least the components near to the high sensitivity f_s frequency are recorded with good quality.

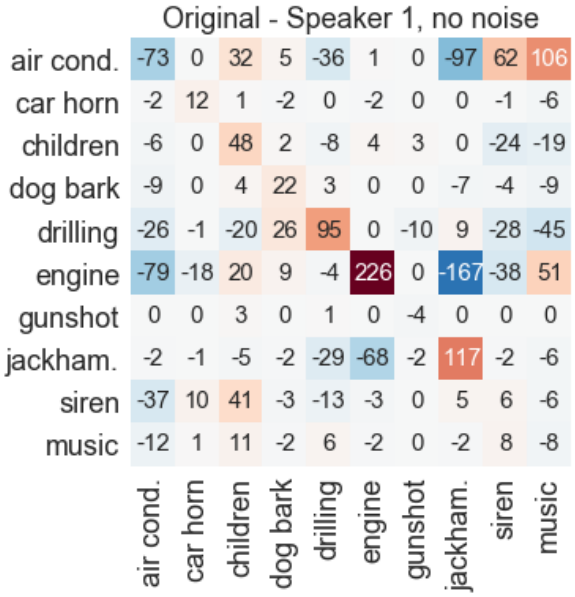


Figure 3.7: *The results of the baseline classifier trained on the original dataset are compared to the results of a classifier trained on a transformed dataset by illustrating the difference between their confusion matrices.*

3.4.2 Experiments and potential applications

The previously presented simple AED scenario can be easily extended to more complex cases. For example, when a building or an area contains multiple speakers that are inter-connected and are driven by a central driving source, the whole area can be protected by deploying a single device, which is listening on the driving line. Similar setups can be found in hospitals, stations or schools, where short statements are occasionally announced through the speakers, but between these active periods, the inactive speakers could be used for event detection. If an event takes place in the vicinity of one of the speakers, the generated reverse mode signal is being transmitted on the driving line. Since the driving lines are connected, all of the speakers produce reverse mode signals on the same line, thus one listening device is enough to protect the entire area. The location information is lost but the deployment becomes effortless. This scenario is presented in Figure 3.8a, where an experiment is illustrated with two speakers placed in the same room but separated by three meters and rotated to opposite directions. The speakers' driving lines are connected. During the experiment, four claps were produced at different positions indicated in Figure 3.8a: the first clap happened close to the speaker 'L', the second and third events were closer to the speaker 'R', and the last one happened near to speaker 'L' again. The recorded reverse mode signal is presented on the right side of Figure 3.8a, where the four events can be easily located. When an event happens close to any of

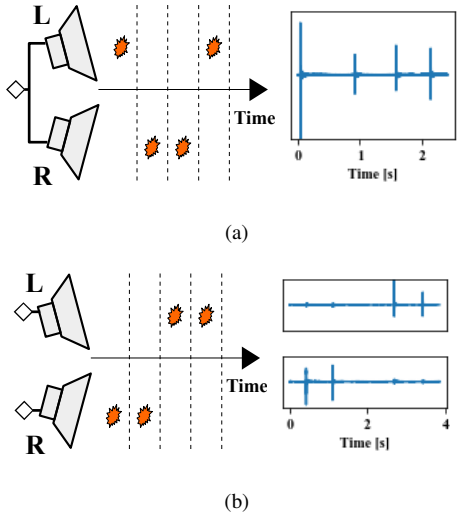


Figure 3.8: *Possible configurations of reverse mode based event detection setups. In (a), the speakers are connected, thus both of them generate their reverse mode signals on the same driving line. During the experiment presented in (a), four claps were produced with different relative locations to the speakers. The claps are all detectable in the reverse mode signal shown on the right side of the figure. In (b), a similar experiment is presented but the speakers have their own dedicated driving line.*

the speakers, the single monitoring device can detect it on the driving line.

Contrary, if there are multiple speakers with given positions and individual driving lines, the location of an audio event can be estimated by listening on all of the driving channels. A coarse localization relies only on the amplitudes of the recorded signals. The location of the speaker with the maximal reverse mode signal amplitude corresponds to the location of the event. An experiment presenting this idea is shown in Figure 3.8b, where the setup is similar to the previous one in Figure 3.8a. Two events were close to the speaker 'R' and two were close to the speaker 'L', which location information can be derived from the reverse mode signals presented on the right side of the figure. The method could be extended further by applying three or more speakers. In that case, a 3D location of the event could be estimated from the time difference of arrivals (TDOA) measurements.

The utilization of reverse mode speakers could be interesting in areas, where already deployed speakers are available and with minimal hardware changes, new applications would become achievable. For example, all smartphones have at least two loudspeakers and one microphone, which configuration is presented in Figure 3.9. In these highly integrated devices, the introduction of new hardware parts may result in huge costs, but the already included speakers' utilization could be maximized by extending the capabilities to record reverse mode signals from them. With this modification, smartphones would have three microphones, thus complex audio localization

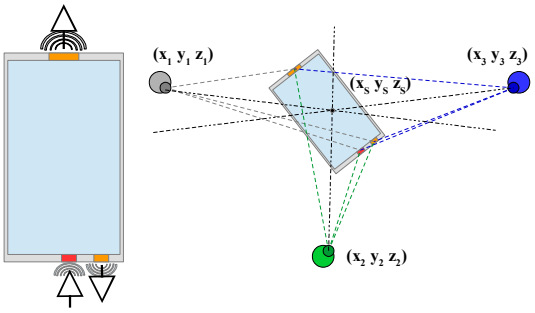


Figure 3.9: *Possible utilization aspect of the loudspeakers in smartphones. They contain at least two loudspeakers and one microphone, which configuration is illustrated on the left side of the figure. On the right side, an example is shown: by converting the speakers to microphones, the indoor localization of the phones would become possible with the help of acoustical signals.*

applications would become possible. One such application is the localization of the incoming sounds, which could lead to the coarse localization of the user relative to the phone. Another potential application is noise reduction during calls. The noise model could be measured more accurately with source separation, which becomes easier with three input channels. Not only the user but the device itself could be localized too, which is a challenging task indoors. This scenario is presented on the right side in Figure 3.9, where a smartphone records the output of three dedicated sound sources with fixed locations. Based on the TDOA measurements, the smartphone could localize itself relative to the base stations. The communication could take place above the human audible range with the help of short chirp signals.

3.4.3 A physical implementation

Reverse mode speakers are capable of recording audio events, thus it is reasonable to utilize them in event detection scenarios. These possibilities were mentioned in the previous sections. Here, one realization of such a device is presented, called the *Smart Speaker*, which activates event detection during the inactive periods of speakers.

The Smart Speaker device provides the normal, direct mode functionality for the user, but between the active sound-radiating periods, it turns the speaker into reverse mode. This is done by controlling two relays to switch between the two directions. Relays are applied instead of transistors because the driving line transmits AC current and the signals' amplitude-ranges are unbounded. The schematic of the proposed device is presented in Figure 3.10. The speaker is connected to the driving source (*D.S.*) in direct mode. Contrary, in reverse mode, the speaker terminals are connected to the input of an instrumentation amplifier, which offers high gain and high common-

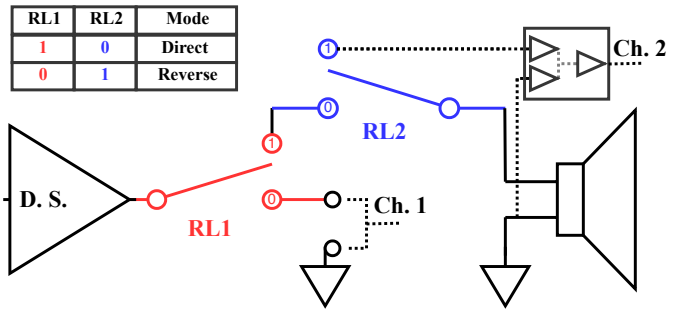


Figure 3.10: *A possible realization of the Smart Speaker device. It alters the states of the relays to switch between direct and reverse modes.*

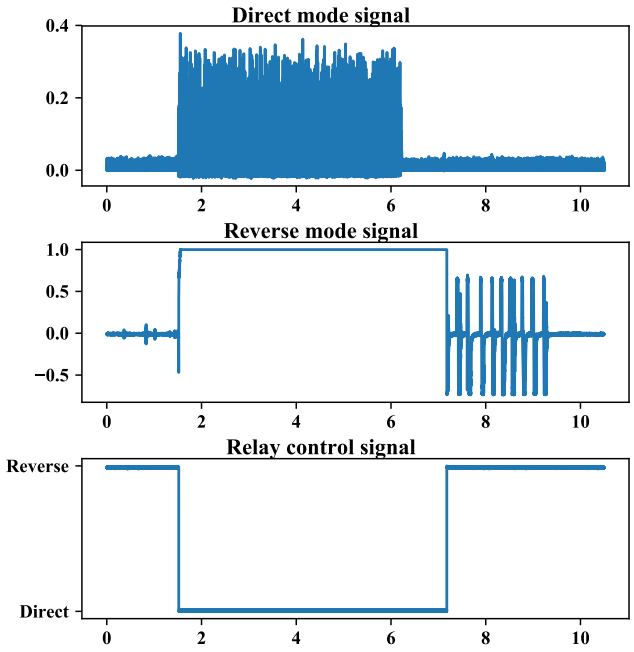


Figure 3.11: *An example recording illustrating the working mechanism of the Smart Speaker device. When the driving source becomes active, the device immediately triggers the direct mode state. After the driving source becomes and remains inactive, the device reactivates the reverse mode state. This switching mechanism based on the direct mode signal can be observed.*

mode noise rejection. The output of this amplifier is connected to an analog-to-digital converter (*Channel 2*) to record the reverse mode signal. In this state, another ADC channel (*Channel 1*) is used for the monitoring of the driving source, thus in case of activity, the direct mode state can be triggered quickly (without audible information loss).

The proposed mechanism was implemented in a pilot setup with the help of a

microcontroller and additional hardware components. The functionality of the device is presented through a real-world recording shown in Figure 3.11. When the speaker is in reverse mode and the driving source becomes active, the device immediately changes the state from reverse to direct mode, thus no audible information is lost. If the driving source becomes and remains inactive for a given time interval (1 second in this case), the device reactivates the reverse mode. This switching mechanism based on the driving source activity can be followed in Figure 3.11.

The Smart Speaker device is simple and easy to install. The jack that was previously plugged in to the driving source needs to be connected to the Smart Speaker and only an extra cable is needed to hook up the device and the driving source. The modification is minimal and the speaker system is extended with a new functionality. To provide a fully-functional solution, one needs to take into consideration the power supply and communication interface of the Smart Speaker device, but these requirements can be solved with off-the-shelf components.

Level of action

The task of the Smart Speaker device is to collect acoustic events. The processing of these captured sounds can happen in different ways taking into consideration various requirements including power consumption, complexity, privacy, accuracy, etc. To present feasible options, three such scenarios are examined in this subsection. These simulation-based experiments evaluate well-known AED setups and their sensitivity to the information loss and increased noise level introduced by reverse mode speakers.

As earlier, traditional audio datasets were used during the experiments, which were transformed by reverse mode simulations. Here, the combination of three different audio datasets was used. The first one was the urban sound dataset [75] introduced earlier. The second dataset [30] contained suspicious events: *glass breaking*, *gunshots*, and *screaming*. The third dataset was collected for speech recognition [62]. Here, only a small portion of the dataset was used and only the sound content was processed.

The sound files had various lengths, sample rates, bit-depths and number of channels. To unify these parameters, all of the recordings were converted to a single channel, 16-bit format with a sample rate of 22050 Hz. From the recordings, one second long excerpts were extracted in a way that the interesting parts of the events were placed at the beginnings of the extracted sections. To generate the reverse mode transformed datasets, each recording went through the reverse mode simulation. A full-range speaker's behaviour was simulated (Speaker 1 [22]).

Three different scenarios were examined. These scenarios require different strategies to provide the proposed functionalities and the reduction in signal quality caused by the reverse mode affects these methods differently. To investigate this negative ef-

fect in the proposed scenarios, the detection algorithms were run both on the original and on the transformed datasets as well and the accuracies were compared.

Scenario I. - `Streaming`: In this setup the presented Smart Speaker device is extended with a CODEC chip, thus real-time audio signal compression can be achieved. This compressed reverse mode signal is forwarded to a base station through wireless communication, where complex analysis of the signal can happen. This scenario should offer the highest detection accuracy since sophisticated, neural network based methods can be run on the server. However, the power consumption is also increased and the privacy is violated as all the events, including conversations, are sent to a server.

In this setup, a convolutional neural network was trained on the original and transformed datasets separately. The neural network structure was similar to [74], and the input data had a format described in [65]. The algorithm was trained to detect *gunshots, glass breaking*, and *screaming* separately and to reject any other noises (all the other classes). These experiments were similar to the ones presented earlier in Section 3.4.1. To avoid model overfitting, the dataset was split into training, validation and test segments (70%, 15%, and 15% respectively). The validation loss curves were analyzed during the training process and the learning rates were optimized based on these investigations.

On the original dataset, the classifier achieved high, 96% accuracy. As the nature of the examined events were mainly different, the classifier could separate them well. On the reverse mode transformed dataset, the classifier obtained 93% accuracy. As the higher frequencies were attenuated by the speaker, information loss occurred, which led to more miss-classifications. Figure 3.12 illustrates the confusion matrices

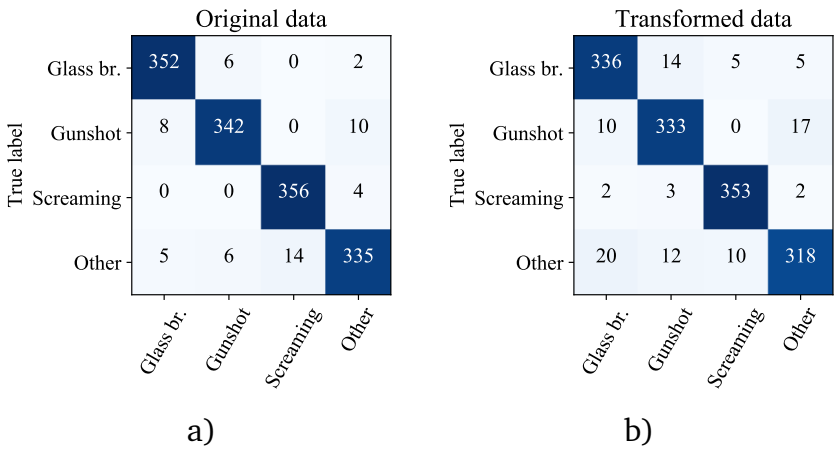


Figure 3.12: *Confusion matrices of the CNN classifiers on the original (a) and transformed (b) datasets in Scenario I.*

of the trained classifiers, where the source of the accuracy reduction is observable. In accordance with the previous simulation based results, the performance on the transformed dataset dropped but not significantly.

Scenario II. - `Local Filtering`: In this scenario, the device sends only the 'interesting' events to the server. This requires local decision making about the recorded sound excerpts. As this filtering should run in nearly real-time, only simple algorithms can be used. In this case, the one second long signals were split into five segments, and from each segment, the zero-crossing-rate, the signal energy, and the maximal amplitude were extracted. That resulted in 15 features per recording. A decision tree classifier was trained based on these features. This method offers a simple and fast prediction even on an embedded system. The 'interesting' events came from the following classes: *gunshots, glass breaking, screaming, dog bark, car horn, siren*. All the other classes were considered as 'not interesting', including the *speech* recordings.

The trained classifiers achieved 87% and 85% accuracies on the original and transformed datasets, respectively. 30% of the data was used for testing. The decision trees were pruned and the hyper-parameters were optimized to avoid overfitting, which is a common problem with these classifiers. Table 3.2 presents the details about the performances of the trained classifiers. The exact classification of the events was not a requirement and as the nature of the events are well-preserved also by the reverse mode speakers, the difference between the two classifiers is marginal.

Scenario III. - `Local Classification`: In this scenario, the simple, microcontroller-based solution is replaced by a more complex digital signal processor (DSP), which offers enhanced code execution capabilities and memory sizes. The recorded reverse mode signals are not forwarded, instead, detection algorithms are operated locally. The complexity of these methods is still way below compared to modern neural networks. 23 easy-to-compute features were extracted that mainly represented the frequency content of the recorded signals [87], e.g. spectral roll-off, spectral centroid, spectral flatness, spectral bandwidth, MFCC features, etc. The dataset was organized similarly as in Scenario I. Two types of classifiers were used. A decision tree (*DcsTr*) and a Multi-layer Perceptron (*MLP*) were trained on the original and transformed datasets separately.

Table 3.2: *Performance of the decision trees trained to detect 'interesting' events in Scenario II.*

	Accuracy	**Recall**	**F1 score**
Original	87%	88%	88%
Transformed	85%	87%	87%

Table 3.3: *Class-wise accuracies in the four tested cases in Scenario III.*

	Glass br.	Gunshot	Screaming	Other
Original: DcsTr	83%	88%	91%	92%
Speaker: DcsTr	75%	88%	89%	87%
Original: MLP	91%	87%	85%	98%
Speaker: MLP	83%	98%	65%	99%

The decision trees' parameters were optimized and the depths of the trees were restricted to avoid overfitting. The MLPs had two hidden layers. The learning rates were changed adaptively and the performances of the classifiers were analyzed and optimized through the evolution of the validation score. Table 3.3 presents the class-wise accuracies in all the four tested cases. The test dataset contained 1440 samples, 360 samples per class.

Summary: From the presented results, it can be concluded that all three Scenarios could be applied. In Scenario I, the classification accuracy is maximal, however, a constant power supply and a reliable communication interface are required. Furthermore, the security of data transmission and storage should be carefully designed to prevent possible cyber-attacks.

Scenario II balances between task offloading and local processing. The detection algorithm is very simple and increased noise level could easily mislead the method. Furthermore, false negative samples will never be sent to the base station, therefore, the overall detection accuracy is less or equal than in Scenario I.

In the third Scenario, the device protects privacy as the recorded raw signal is not transmitted. The trained classifiers achieved acceptable accuracies, however, extra noise might be able to diminish the slight differences in the spectral-based features, therefore, these simple classifiers would perform poorly. This drawback is better handled by convolutional networks, so Scenario I offers more robustness.

In all the tested situations, the reverse mode speakers had a slightly negative impact on the detection accuracies. This is reasonable if the nature of the reverse mode is taken into consideration. Yet, the decrease confirms that these setups should only be used when a completely deployed system of speakers is already available and a quick, temporary solution is needed. The choice between the scenarios should be based on the required accuracy, privacy, and available resources.

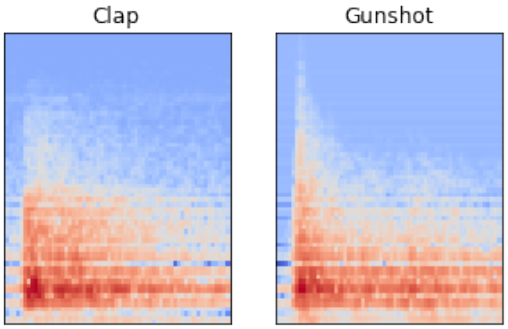


Figure 3.13: *Two example input mel-spectrograms. The clap was recorded through a reverse mode speaker, while the gunshot comes from reverse mode simulation.*

3.4.4 Clap detector

In the previous sections I have presented results originated from simulations. In the current subsection a clap detector is introduced that is based on real-world clap recordings that were captured by a reverse mode speaker with the help of the Smart Speaker device.

Data collection:

The construction of a deep learning based clap detector requires the collection of relevant, high-quality data. In this proof-of-concept scenario, the detector distinguishes six separate classes: *car horn, dog bark, jackhammer, gunshot, siren, clap*. All of the events have an impulsive nature that makes the problem more challenging.

The claps were recorded with the Smart Speaker device, which was connected to a 2” full-range speaker. The other events came from the dataset containing environmental sounds from urban environments [75]. These events went through reverse mode simulations to convert them into forms as they would have been recorded by the 2” loudspeaker. We have already seen that these simulations produce realistic output, thus the real-world recordings and the simulated ones are comparable. This similarity between two such events, a clap, and a gunshot, is observable in Figure 3.13.

The claps were recorded at a sampling rate of 44100 Hz. The other recordings were resampled at this frequency to unify this parameter. An 8192-samples long part (180 ms) was taken from all of the recordings that contained the interesting, impulsive sections. Then, the spectrograms of these short signals were calculated with 512-point FFTs and a hop-length of 128. The resulting spectrograms, excluding the claps, went through image-based reverse mode simulation explained in Section 3.3. These reverse mode spectrograms were converted to Mel-scaled spectrograms with

64 bands. These served as the input for the classifier algorithm. The final input size was 65×64. Example data samples can be observed in Figure 3.13. The final dataset contained: 446 recorded claps, 429 car horn, 1000 dog bark, 374 gunshot, 1000 jackhammer, and 929 siren events. The imbalanced nature of the dataset was handled during the training process by adjusting the class weights. 70% of the data was used for training, 20% for testing and 10% for validation.

Classification:

In the previous section, different Smart Speaker operation strategies were investigated. Here, the Local Filtering scenario was assumed, in which the Smart Speaker device implements some sort of filtering, thus it only transmits interesting events to a base station that has more computational power to run complex classification algorithms.

It was already mentioned that in sound classification tasks, state-of-the-art results have been achieved by using convolutional neural networks on spectrogram 'images'. These structures interpret their input as images and extract high-level features that are being evolved during the training process. An extended version of this structure with additional recurrent layers was trained to perform the clap detection task. The model was published in [25] and originally classified spoken commands. The advantage of this model is the low number of parameters, which is around 193,000. This is at least a magnitude smaller than the other CNN models with comparable accuracies [13, 65, 74]. First, the model applies a set of convolutional layers on the mel-spectrogram to extract local features, which features are fed into long-short term memory (LSTM) units to capture two-way long term dependencies. In the final stage, classification labels were produced by three fully-connected layers. Further details about the architecture can be found in [25]. The input and output layers of the original model were adjusted to the current input shape and prediction output format. The model and the training process were implemented in Keras [16]. Categorical cross-entropy served as the loss function, the mini-batch size was 8. With Early-Stopping based on the validation accuracy the training converged after 20 epochs. The source code of the training process can be found in [37].

Results:

The trained classifier's performance is detailed in Table 3.4 and in Figure 3.14. The class-wise measures show that the claps were reliably recognized. The overall accuracy of the classifier was 91%. From the confusion matrix, the source of the error can be localized, which includes the *Car horn*, *Dog bark*, and *Siren* classes. The two best-performing classes were the *Gunshot* and *Claps* classes, which samples had very

Table 3.4: *Detailed results of the trained classifier.*

	Precision	Recall	F1-score	Support
Car horn	0.93	0.80	0.86	88
Dog bark	0.91	0.85	0.88	203
Gunshot	0.95	0.95	0.95	73
Jackhammer	0.86	0.99	0.92	179
Siren	0.90	0.92	0.91	190
Claps	1.00	0.97	0.99	103
Accuracy			0.91	836

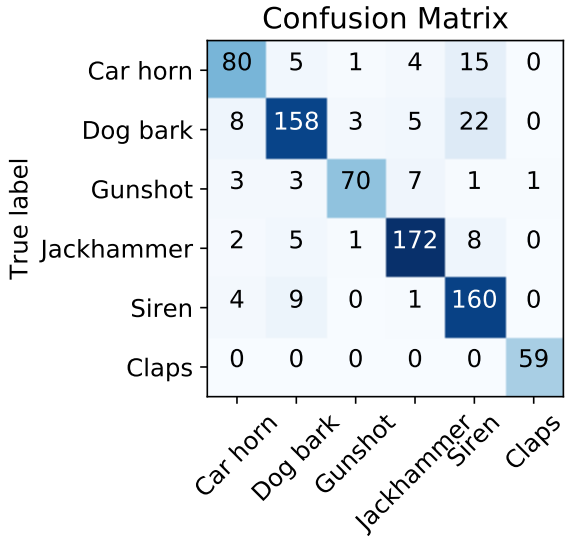


Figure 3.14: *Confusion matrix of the trained classifier. The source of the relevant error can be localized, which includes the Car horn, Dog bark, and Siren classes.*

similar spectrograms (as was shown in Figure 3.13), but remained diverse enough to guide the classifier.

This section presented utilization aspects of reverse mode speakers with simple experiments included. In the next section, a more challenging scenario is analyzed, where the reverse mode event detection remains the task while the speaker is being actively driven by the driving source.

3.5 Active reverse mode

As was discussed in the previous sections, reverse mode speakers are able to record sound with fair quality that allows the implementation of acoustic event detection applications. In those scenarios, the speakers were inactive, meaning that their driving source was inactive during the recordings and the idea was to utilize these idle periods. Contrary, in the current section *active reverse mode* scenarios are analyzed, where acoustic event detection is required while the speakers are being actively driven in parallel with the acoustic events. These situations are typical in public spaces (stores, cafes, restaurants, shopping centers, etc.), where low volume music is played constantly from distributedly deployed loudspeakers. These places may become the targets of violence against the public or terror attacks. Fast responses to such events could save many lives.

During the direct mode, the driving signal forces the cone to oscillate. The sound radiation efficiency of speakers is below 10%, meaning that only a small portion of the electrical power is converted to acoustic power. Therefore, the driving line must carry high energy electrical signals to produce sound in the human audible range ($>$0 dB$_{SPL}$). The convenient range, e.g. for music listening, is around 40-60 dB$_{SPL}$. To generate these levels, the driving amplifiers must output signals with amplitudes starting from several volts.

During the *active reverse mode,* the speaker is actively producing sound while external acoustic signals are also reaching the cone. It was already demonstrated that external incoming sound waves generate reverse mode signals on the driving line. This effect remains the same even when the driving source is active and in this case, the sum of the direct and reverse mode electrical signals is present on the driving line. This behavior is explained by the superposition of the direct and reverse mode mechanical forces that act at the same time on the diaphragm. This simultaneous action of the forces is modeled in the equivalent mechanical circuit by connecting two force sources in series, as shown in Figure 3.15. The $F_{reverse}$ source represents the reverse mode force produced by the pressure waves and the F_{direct} source represents the direct mode induced force. It can be observed from the model

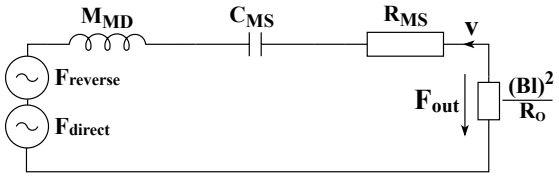


Figure 3.15: *The equivalent mechanical circuit of the active reverse mode. The previously showed circuit was extended by another force source that represents the forces generated by the driving signals.*

that the superposition of forces leads to the superposition of electrical signals, thus an observer device connected to the driving line has access only to the superposition of the direct and reverse mode signals. Let us denote this combined signal with **y**, which is the sum of the reverse mode signal **r** and the direct mode signal **d**.

The active reverse mode event detection is challenging, which originates from the amplitude difference between the direct and reverse mode signals. As was illustrated earlier, reverse mode sensitivity is low and the generated signal is in the range of several millivolts. These little changes in signals with several volts of amplitudes should be detected to perform the event detection task. In the following, several detection approaches will be presented that offer ways to separate the **r** and **d** signals from **y**.

Loud events: In those cases when extremely loud events like gunshots or explosions should be detected and the radiated audio volume is low, the amplitude of signal **r** is close to the amplitude range of the signal **d**. Thus, **r** is a significant portion of the resulting signal **y** and its pattern is distinguishable in the spectrogram. Neural networks are capable of detecting such patterns even with complex background noises. This capability is often utilized as a data augmentation method, or during training, called between-class learning [84]. Therefore, classifiers could be trained to detect and even to classify active reverse mode events directly from the signal **y**.

Normal events: When the amplitude of the reverse mode signal **r** is much smaller than the amplitude of the driving signal **d**, **r** cannot be directly detected from their sum **y=d+r**, thus events with normal SPL levels cannot be recognized. With similar radiated and incoming SPL levels, the generated reverse mode signal **r** is at least $100\times$ smaller than the driving signal **d**. In these cases, more information is needed to extract the **r** from **y**. Two simple configurations that provide this extra knowledge are presented in Figure 3.16. These setups give access to the original version of the signal **d**, thus with a subtraction, **r** can be reconstructed from **y**: **r=y-d**. In both configurations, the stereo output must be mixed down to a single mono channel and the two speakers should be driven by the same signal. In Figure 3.16a, the *Left* channel drives the speakers and this channel is affected by reverse mode signals. In parallel, the *Right* channel is used to acquire the original version of **d**. From **y** and from the original **d**, the reverse mode signal **r** can be calculated by simple subtraction.

In Figure 3.16b, a similar configuration is shown, but both channels are actively used. If the event happens closer to one of the speakers, or only one of them records it, the reverse mode signal can be reconstructed by taking the difference between the two channels. If both speakers capture the same event, but with different relative distances from it, the times of arrivals and the amplitudes of the recorded reverse mode signals will be different, thus detection based on their difference remains possible. An experiment with this setup is presented in Figure 3.17, where the word

'Hello' with 72 dB_{SPL} was pronounced 1 meter from the *Right* speaker, while both loudspeakers were radiating a pure sine wave with 440 Hz at 78 dB_{SPL}. The *Left* speaker didn't hear the event. In the upper part of Figure 3.17, a short section of the two recorded **y** signals can be compared. It is observable that the incoming waves have only a very small effect on the *Right* channel. Below this comparison, the difference signal of the two channels, i.e. the reconstructed **r** is shown, where detectable changes above the noise level are visible. At the bottom of Figure 3.17, the spectrogram of this difference signal **r** is presented, which could be the input of a classifier or a speech recognition algorithm.

As was shown, the key factor during the active reverse mode is the separation of the reverse mode signal from the direct mode signal. Here, only direct methods were presented that guarantee this separation by providing the original version of the driving signal. However, more advanced methods could be proposed that would try to predict the value of **y** in a future time point from a short, measured history of the signal. In that case, event detection could be based on the differences between the estimated and real values. This setup would require the measurement of the combined signal **y** only, but has more limitations. The examination of these methods remains future work.

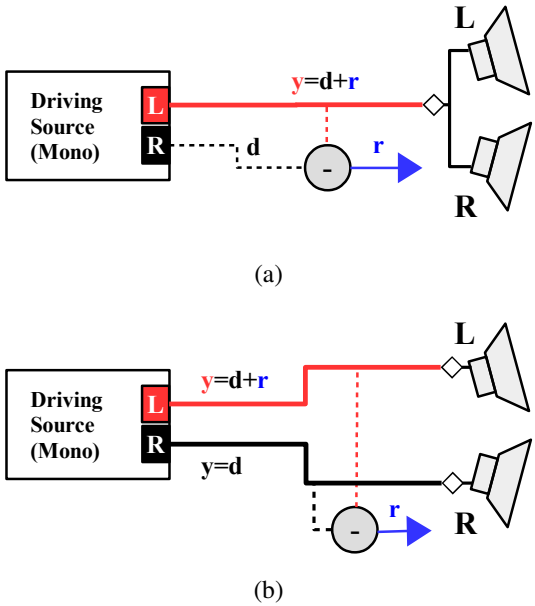


Figure 3.16: *Possible active reverse mode configurations. The goal is to reconstruct* ***r*** *from* ***y****. In (a), only one channel is used to drive the speakers, the other channel serves as the reference signal* ***d****, thus* ***r*** *can be calculated with a simple subtraction. In (b), the idea is similar, but both channels are simultaneously used and their difference produces the signal* ***r****.*

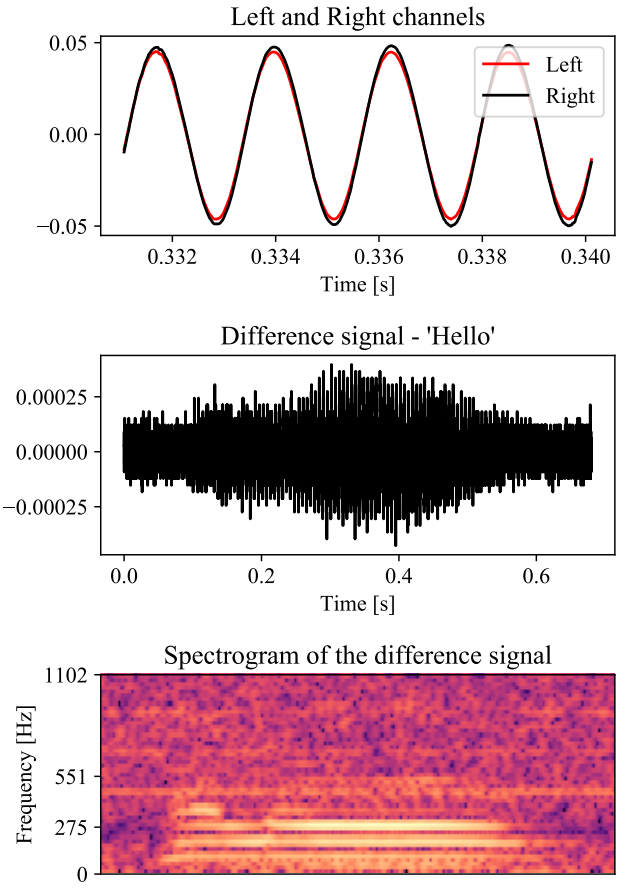


Figure 3.17: *An experiment is illustrated presenting the feasibility of active reverse mode based event detection. The word 'Hello' with 72 dB_{SPL} was pronounced 1 meter from the Right speaker, while both loudspeakers were radiating a pure sine wave with 440 Hz at 78 dB_{SPL}. The Left speaker didn't hear the event, so this channel is used as the reference signal. It can be observed that the word could be reconstructed despite the hardly detectable changes in the Right channel's signal.*

The active reverse mode introduced in this section requires the detection of the reverse mode signal **r** with at least two magnitudes lower amplitude than the driving signal **d**. Preliminary results and ideas were presented that offer ways to acquire and detect acoustic events through actively radiating loudspeakers. As a conclusion, it can be stated that the event detection is possible with restrictions on the speaker configurations and event types. For example, gunshots or explosions could be detected reliably if the system had access to the original form of the driving signal.

3.6 Discussion and concluding remarks

This chapter investigated the behavior of the 'microphone' mode - referred to as the *reverse mode* - of loudspeakers. In this state, the speaker converts sound to electrical signal, thus its environment becomes observable. The work introduced the analysis of the reverse mode through the formation of an equivalent mechanical circuit and included the results of experiments, simulations, and measurements. Possible utilization perspectives were also proposed and a proof-of-concept data-driven device that implements clap detection was explained. During these investigations, it was assumed that the loudspeaker was inactive. However, the chapter also included the analysis of the more challenging *active reverse mode*, in which acoustic events are being detected through the speaker while it is also actively radiating sound.

The analysis of the reverse mode behavior showed that an increased sensitivity zone exists around the mechanical resonance frequency. This frequency interval is recorded with good quality, thus events with relevant frequency content in this region can be detected accurately. The resonance frequencies of widespread full-range speakers overlap with the human voice fundamental frequency range, which allows speech to be captured with acceptable quality. This coincidence may raise privacy concerns and supports the rumors about the application of speakers in spying attacks. Therefore, privacy should be taken into consideration during the design of such event detection systems.

As was presented in Section 3.4, the inactive periods of speakers could be utilized to implement acoustic event detection. An example device called 'Smart Speaker', and a classifier distinguishing impulsive events were explained. It showed that with the combination of a simple embedded device and a neural network model, claps can be accurately detected through a reverse mode speaker. The main advantage of this setup is that with minimal deployment effort and hardware extension, the existing loudspeaker systems could be turned into complex security systems.

In Section 3.5 a more challenging and more general aspect of the event detection is presented, where the acoustic events are being detected during the active periods of loudspeakers. It requires the separation of the dominating driving signal and the almost undetectable, weak reverse mode signal. This can be achieved in slightly modified speaker configurations (mono channel instead of stereo) and only events with high SPL levels can be detected. Despite these restrictions, the deployment is simple and the system would offer real-time reactions to crimes involving guns or explosives. The real-world test of the idea remains future work.

3.7 Contributions

The author of this PhD thesis is responsible for all the contributions presented in this chapter, which include the following main novelties:

II/1. I proposed the idea of using loudspeakers for audio event detection by employing their reverse mode functionality. I carried out the theoretical modeling and analysis of the reverse mode, and supported it by real experiments.

II/2. I investigated through simulation experiments the reverse mode speakers in acoustic event detection scenarios.

II/3. I designed and implemented an audio event detector device based on the reverse mode functionality, and demonstrated its use by a simple, data-driven clap detector.

II/4. I investigated the loudspeakers' active reverse mode through theoretical modeling and analysis, and some experiments, when the speakers may be used for acoustic event detection while they are actively radiating sound.

Chapter 4

Automated Pupillometry

In this chapter an image and video processing application is presented that involves the analysis of videos recorded during animal experiments. These medical investigations examine the effects of schizophrenia-like symptoms on the nervous system of rats by extracting objective measures from their pupillary light reflex (PLR) curves. The task is to detect and measure the size of the pupil in consecutive video frames and to analyse the dynamics of the resulting pupillograms. The chapter presents the evolution of the research that includes the development of a traditional image processing algorithm, the hardware-based extension of the experimentation setup, and the development of a data-driven CNN-based method. Medical-related results are also summarized that revealed significant differences between the PLR curves of the healthy and test animals, thus supporting the understanding of the schizophrenia-related pathophysiological disorders in the examined rat model.

4.1 Introduction

Non-invasive techniques and their usage in diagnostics are active research fields in medical science. Most of the well-known solutions rely on computer-aided imaging, such as MRI, CT, SPECT, and PET, etc. One of the simpler techniques is pupillometry, which can be used to analyze mental diseases and their pathophysiological changes involving the pupillary reflex. Fundamentally, pupillometry is the measurement of the pupil diameter during and after visible light stimuli, which induce pupillary light reflex (PLR), i.e. the contraction of pupil in response to light. PLR-derived metrics can be used to monitor the extent of neurological diseases and the patient's response to therapy. In this work, schizophrenia is investigated through pupillometry experimentation. Earlier studies examined the parasympathetic function in the context of schizophrenia through the use of pupillometry. These investigations pointed out that the pupil diameter is larger and the reflex is slower in patients with schizophrenia.

The concept of pupillometry is to record the pupillary reflex with a camera and af-

ter the experiments, these videos are processed offline. Measuring the diameter manually in each frame with the help of a simple annotation software is time-consuming and non-reproducible, especially in cases, when the number of videos is high. Automated computer algorithms may provide faster, more reliable, and reproducible solutions to the analyzing process. These methods usually use the prior information about the dark color and circular shape of the pupil. From an image processing viewpoint, the problem can be traced back to circle or ellipse detection and the combination of well-known circle detection methods with special, pupillometry-related information may result in accurate and robust solutions.

In this chapter, the automation of pupillometry of rats is detailed in two steps. In the first part, a classical image processing method for pupil diameter measurements is presented that achieved high accuracy, but many videos were left out as the experimentation setup produced videos with poor quality. This setup was reorganized to overcome the quality drawbacks. A small hardware extension on the recording camera drastically improved the measurement process and more videos could be recorded with high quality, which led to the development of a data-driven pupil detection method. These improvements are detailed in the second part of the chapter.

4.2 Pupillometry with classical methods

The presented research focuses on measuring the pupil diameter of rats in infrared (IR) video frames. These measurements support a medical research that investigates the influence of schizophrenia on pupillary light reflex. Developing reliable and predictive animal models for any complex psychiatric diseases, such as schizophrenia, is essential to understand the neurobiological basis of the disorder. Recently, a new, selectively bred rat model of schizophrenia has been developed, named WISKET [40, 41, 46, 64]. Pupillometry-based clinical studies in schizophrenic patients revealed impaired autonomic regulation [4, 5, 38, 53, 96]. The PLR was also investigated in rodents [60], however, only our study provides data from animal models of schizophrenia. In summary, the goal is to reveal objective measures of the disease through animal experiments that might be examined later in clinical research.

The pupillometry experiments took place in a dark room, therefore, the rats' pupils were more dilated. The room was illuminated by an intense infrared light, which is invisible for the rats, thus information about their pupil could be collected by an IR camera without negative influence. During the recorded part of the experiments, rats were held by hand on a desk while a short visible light impulse was emitted into their eyes, which induced the PLR reaction. The camera was placed close to the eye before the experiment to record the pupil response to the light stimulus.

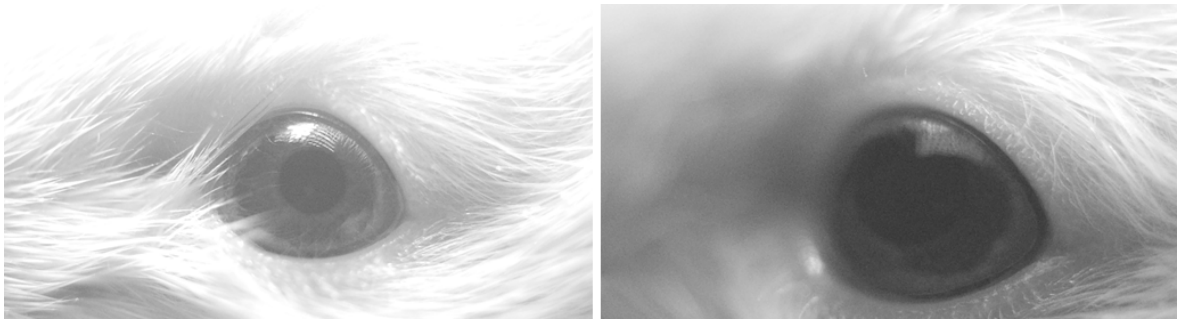


Figure 4.1: *Example frames from the pupillometry videos. A good quality image is presented on the left side and a bad quality image on the right side.*

The videos had quality drawbacks. The breathing and slight movements of rats caused significant blur, which makes the segmentation of the pupil impossible with simple binarization. Beyond these artifacts, the rats were albinos and had red eyes that reduced the contrast between the pupil and the iris regions. Low lighting conditions also forced a higher ISO level value, which implies noisier images. The reflections of the illuminating IR LEDs on the glossy eye were obscuring a big part of the pupil boundary, while other overlying entities such as the whiskers made the pupil many times nearly undetectable. In Figure 4.1, two example images are shown. On the left side, the dark region in the middle of the eye is the pupil region, which is easy-to-detect. Contrary, on the right side, the image is blurred, has a low contrast and the pupil region is barely distinguishable from the iris region.

4.2.1 Related works

As was mentioned, pupil detection can be interpreted as a circle or ellipse detection problem. In this subsection, related traditional methods are summarized. One of the popular approaches is based on the Hough transform. The Circle Hough Transform (CHT) detects circles with a given radius. It uses an accumulator space and local maxima searching. CHT works on edge images and for each edge point, it increments the accumulator values that correspond to points located in a given distance from the selected edge point. The accumulator space is a 3D space. Two dimensions correspond to the coordinates of the center of the circles, and the third to the radius. After processing all edge points, local maxima appear at the accumulator points that correspond to the circle centers. Modifications to the CHT (edge orientations, single accumulator space for multiple radii, randomization, etc.) have been introduced to reduce the computational cost and increase the detection rate. For more details see [12, 43, 95]. In our case, the CHT cannot be applied directly because there may be a substantial blur, the relevant part of the pupil boundary may not be observable in the edge image, and the pupil shape is usually not a perfect circle.

Another idea is based on the fact that the image gradients at the circle boundary point outwards from the center of the circle (with dark circle, white background). The checking of some simple properties for all gradient vector pairs leads to a fast circle detection algorithm described in [69]. The first step is to calculate the gradient image and because of the symmetry of the circle, for each gradient vector there will be a pair-vector in the opposite direction. For a given vector **V**, its pair vector **W** satisfies two basic conditions: the absolute difference between the two directions should be nearly 180 degrees and the angle between the vector **V** and the line connecting the bases of vectors **V** and **W** should be nearly 0 degree. In the next phase, the corresponding vector pairs are formed. Then, candidate circles are computed for all vector pairs. In the last phase, suitable circles are extracted from the candidates. In our case, the reflections and blurry boundaries do not generate a consistent gradient vector space, which prevents the detection of the circular shaped pupil.

Optimization-based methods are also popular. These methods use well-known optimization techniques to find local extrema in an appropriately chosen function space. Some of them inspired by biological notions. Such approaches are the Genetic Algorithms (GA) [3], Bacterial Foraging Algorithm Optimizer (BFAO) [21], Harmony Search Optimization [17], Artificial Bee Colony (ABC) optimization [19], and several others [18, 20].

Pupillometry is a long-known method that is used to objectively characterize the pupil's response to light stimuli. Related problems are eye tracking, pupil center detection, and pupil diameter measuring and all of these methods are part of an automated pupillometry application. Earlier such simple pupillometry methods tried to determine a suitable threshold value and binarize the eye image, which is later processed by shape analysis or ellipse fitting algorithms [15, 47, 51, 59, 70]. In [26], the authors proposed a fully-automated procedure for pupil segmentation based on the level set theory. The level set formulation is able to deal with the complex topology of biomedical images regardless of the initial level set configuration. They used 4-level segmentation, which was suitable to extract the pupil from the eye image, and measured various morphological parameters, such as the pupil's diameter, centroid, and area. In [97], the proposed algorithm used the curvature characteristics of the pupil boundary to determine the visible portion of the pupil, which provided improved estimates of the pupil center point even if it was obscured by a host of artifacts. The curvature algorithm discriminated between edge points that lie on the smooth pupil boundary and those that lie on the intersection of the pupil with eyelids, eyelashes, corneal reflections or shadows. The non-occluded boundary points were used as input to an ellipse-fitting procedure that provided a robust estimate of the pupil center. The above-mentioned papers have described acceptable solutions for the pupil detection problem but these methods have their shortcomings when it comes to handling reflections, blur, and low contrast difference at the same time.

4.2.2 Methods

To overcome the previously mentioned video quality setbacks and accurately detect the pupil, and automatically measure its parameters, a novel ray propagation based method with an energy attenuation model was developed. This algorithm could find the fine contrast changes at the boundary of the pupil, while tolerating noise, reflections, occlusions and blurred images. The proposed technique could also measure the diameter of small pupil with sharp edge as well as big, blurred, noisy ones using the same process.

The procedure seeks to support medical staff in their studies on the connection between schizophrenia and pupillary reflex. Processing hundreds of videos by hand is time-consuming and non-reproducible. To overcome this problem, an automated system was developed that can bring a significant speed-up in the processing stage. The input of this system is a video, which contains the recorded pupillary response to a single stimulus. The output is the pupillogram that expresses the change in the pupil diameter over time. The pipeline of the proposed automated process is shown in Figure 4.2.

Movement filtering

As was mentioned earlier, the animals breathed and made small movements as they have only been slightly sedated and held by hand. These caused rapid motions of the eye in the videos. To compensate this artifact, the recordings were stabilized with the help of the Kanade-Lucas-Tomasi (KLT) point tracker [85]. It was assumed that the size and shape of the eye do not change during one recording (as the rats do not blink) and only translations occur between the consecutive frames. The KLT algorithm tracks feature points throughout the video. Such points are automatically selected by the algorithm. In each frame, the successfully tracked points and their

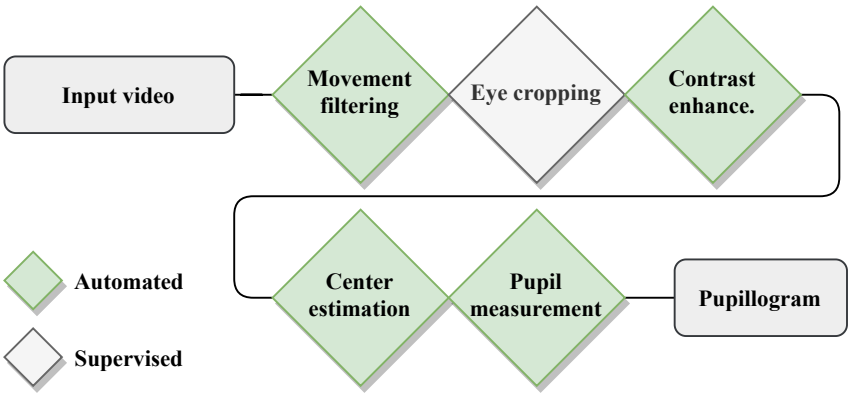


Figure 4.2: *Pipeline of the proposed classical pupil measurement method.*

initial positions are connected by translation vectors. The median value of these translations is used to translate the whole frame back to an estimated initial position. The result of this stage is a video, in which the eye region is stabilized.

Eye cropping and contrast enhancement

The region of interest (ROI) in our case is the eye region, more precisely, the iris region. The diameter of the pupil is expressed relative to the size of the iris, therefore, it is important to determine the iris region accurately. The problem is that the boundary of the eye is not defined exactly because of shadows, hairs, and reflections. Experience is necessary to select the appropriate region that contains only the eye. In the second stage of the pipeline, the algorithm generates an initial ROI guess based on the projections of the dark pixels to the horizontal and vertical axes, but supervision is required. A simple graphical user-interface was developed for the medical experts to assist the selection of the ROI. After they define a bounding box that contains the eye, this region is cropped from every video frame, which is reasonable as we assumed that after the stabilization step the eye's position remains constant.

One of the major problems was that the contrast between the pupil and iris regions was low. A contrast enhancement process tried to increase this difference by assuming that the minimal pixel intensity within the eye region corresponded to the pupil region. By analysing the histogram of the eye's area, it was observable that it was characterized by a bimodal nature (pupil and iris regions). Based on the histogram, an optimal threshold value was estimated that separated the iris and pupil regions. The minimal pixel intensity and this estimated threshold value served as the two main points for a histogram stretching algorithm.

Center point estimation

The proposed diameter-measuring algorithm required a good pupil center point estimation. During this estimation phase the entire video was processed and the output was a single 2D coordinate. This process relied on the novel ray propagation based method that will be described in Section 4.2.3 and the center point estimation will be also discussed in more details in Section 4.2.4.

Pupil measurement

In this step the diameter of the pupil was determined in each video frame. An ellipse was fitted to the filtered boundary points produced by the ray propagation method. More details are given in Section 4.2.3. The output of this process was the pupillogram, a diagram which expresses the change of the pupil diameter during the PLR experiments.

4.2.3 Ray propagation with energy attenuation

As was described earlier, the contrast between the pupil and iris regions was low due to the rats' eyes melanocyte content. Also, the varying relative position between the rats and the light source affected the amount of light reflected back to the camera from the iris. These factors led to significant variance in video quality and made the proper pupil detection challenging. The proposed method could handle these dynamic changes of intensity levels, not only among videos but also frame-by-frame, and remained robust to noise.

In the following, we shall assume that the pupil region has a darker color (lower intensity) than the surrounding iris region and that the cropped video frames contain only one dark circular region, which is enclosed by a brighter region. It does not matter that only a part of the pupil may be visible or it is affected by significant amount of noise and blurring. It is also assumed that the center of the pupil has been accurately estimated and that it remains constant during the recordings.

The proposed method uses notions and ideas taken from the physics of ray propagation. The rays have an initial energy which is gradually absorbed by the medium during the propagation. The amount of energy loss is proportional to the attenuation coefficient of the medium. The medium with no attenuation is called vacuum. Based on these principles, the concept is to radiate rays from a point and use the pixel intensities as the measures of attenuation capabilities. Higher pixel intensity represents a higher attenuation coefficient. The method traces the rays while they travel through the image and uses the information of energy loss characteristics to learn more about the structure of the surrounding regions.

Rays with given initial energies are radiated equiangularly from the estimated pupil center. They are traced until their energy gets completely absorbed and the position of this endpoint ε is recorded. Because of the above-mentioned notions are inherently related to directions and distances from a center point, the 2D pixel coordinates will be determined in a polar coordinate system, whose origin is the estimated center point of the pupil and the polar axis is drawn horizontally and pointing to the right. We denote the angular and the radial coordinates by θ and r, respectively. At point $[\theta, r]$, the pixel intensity is represented by $I_{\theta r}$. For a ray in direction θ, the radial coordinate of its endpoint is denoted by ε_θ.

The basic idea is to treat the pupil region as a vacuum. For each frame, we calculate the vacuum intensity

$$I_v = \frac{1}{|R|} \sum_{x,y \in R} I_{xy}.$$

Here, I_v is defined as the average pixel intensity of a small R region around the estimated pupil center. Then, the original pixel intensities I are shifted to I^* as

follows: $I^* = \max[I - I_v\ , 0]$. If a ray with an energy of E travels through a pixel at $[\theta, r]$ with an intensity of $I^*_{\theta r}$, the ray's energy is updated:

$$E := E - f(I^*_{\theta r}).$$

f is a quadratic weighting function, which emphasizes the low contrast between the pupil and iris regions.

Let us define the set of initial energies $E_I = \{e_k = \alpha + k\delta \mid k \in \mathbb{N}\}$, where α is an initial offset energy and δ is the step size between the different energy levels. It is worth noting that a reasonable upper limit for the energy levels can be determined by considering the intensity values in the image. One possibility might be the measurement of the minimal amount of energy required to reach the border of the selected ROI.

The higher the initial energy the more a ray can penetrate into the high intensity iris region. If the region of the pupil is noise-free, rays with low initial energies can reach the border of the pupil and stop at, or nearly after the boundary points. As the iris region has a higher intensity, these pixels attenuate the energy more and stop the rays more quickly. If there is noise, some of the low energy rays stop before reaching the boundary but most of the higher energy rays still stop close to the boundary.

For a given direction θ, and for each initial energy $e_i \in E_I$, we calculate the corresponding rays' ε^i_θ endpoints. The set of endpoints in a given direction θ is denoted by $S_\theta = \{\varepsilon^i_\theta\}_{i=1,\dots,n}$, where n is the number of rays that have been radiated (the number of distinct initial energy levels). If rays with uniformly incremented initial energies propagate in a homogeneous medium their endpoints will spread equidistantly. However, in our case, the structure of the eye is inhomogeneous, i.e. the pupil region's intensity is close to that of the vacuum, unlike the outer regions that mostly have higher attenuation. This inhomogeneity forms clusters in the endpoints' positions. It can be seen that the casting of rays with monotonically growing initial energy levels leads to endpoint clusters near to the boundary points, specifically, near to the pixels with higher intensities. From here, the task is to properly select the correct endpoint cluster that corresponds to the pupil's boundary. Hierarchical-clustering was utilized to form these clusters from the endpoint positions for a given direction θ based on point-wise distances as follows:

$$c_{\theta j} = \{\varepsilon^i_\theta \mid \text{maximal shortest distance between adjacent endpoints} < d\},$$

where d is a given small constant, $c_{\theta j}$ is a cluster of endpoints, where the distance between two adjacent endpoint is less than d. Earlier, we assumed that beyond the pupil's boundary mostly higher intensities are present. Therefore, it can be seen that rays with growing initial energies end up in the same cluster after reaching the pupil's border as they can only penetrate just a little more and more into the higher intensity

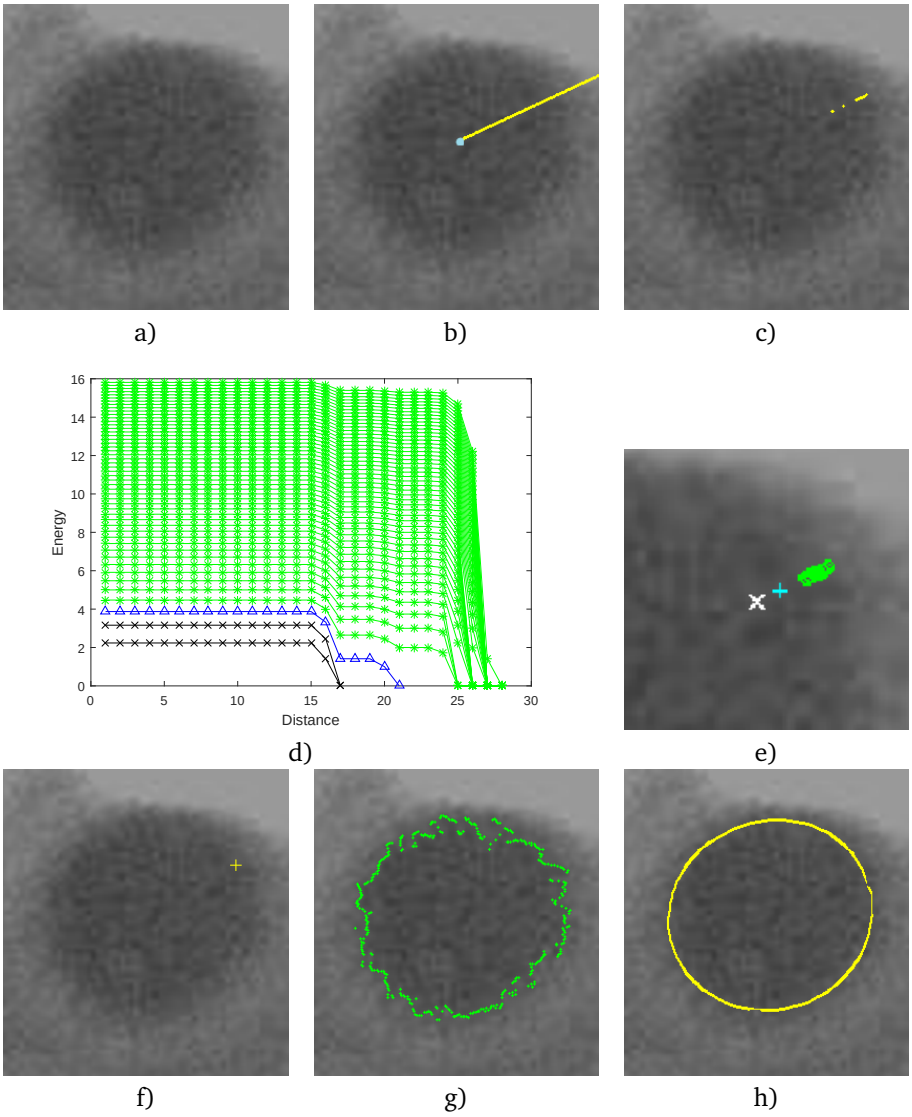


Figure 4.3: *The steps of the proposed ray propagation based method: a) original input image (pupil area); b) selected direction of a ray propagation iteration; c) endpoints of rays; d) diagram of the rays' energy curves with different initial energy levels; e) clustered endpoints; f) selected endpoint in the examined direction; g) selected endpoints for all directions; h) output after the filtering and ellipse fitting steps.*

region. Let us denote $C_\theta = \{c_{\theta j}\}_{j=1,\dots,m}$ the set of clusters for the direction θ, where m is the number of clusters. C_θ is the clustered version of S_θ, so $\bigcup_{j=1}^{m} c_{\theta j} = S_\theta$.

By using a set of initial energies with higher density (i.e. smaller δ), it is also easy to show that the cluster corresponding to the boundary region has the maximal cardinality. So, in the next step of the algorithm, find the cluster $c_{\theta k}$, where k is

$$k = \arg\max_j |c_{\theta j}|,$$

where $|c|$ is the cardinality of cluster c. From the selected cluster, the closest endpoint to the estimated pupil center (i.e. the endpoint with the smallest radial coordinate r) corresponds to the pupil boundary point $\beta_\theta = [\theta, r_{min}]$, where:

$$r_{min} = \min_i \|\varepsilon_\theta^i\| \ , \ \varepsilon_\theta^i \in c_{\theta k}.$$

For all θ directions, collect the selected border points into a set denoted by $\mathbf{B} = \{\beta_\theta\}_\theta$, where θ is a set of directions. In our setup $\theta = \{1, 2, ..., 359, 360\}$. This **B** point-set is the output of the ray propagation based method, which is processed by later stages to estimate the size of the pupil.

Figure 4.3 illustrates the steps of the proposed process. The original image in Figure 4.3a has a good quality with observable contrast between the pupil and iris regions but it is noisy and blurry. For visualization reasons, not the whole eye but only the pupil area is presented. We can follow the above-mentioned steps in a selected direction highlighted in Figure 4.3b. Figure 4.3c illustrates the endpoints of the different rays. In Figure 4.3d the energy-loss curves of the rays with growing initial energies are shown, where the square root of the original values are presented to better highlight lower initial energies. The curves are colored according to their endpoints' cluster membership, which is shown in Figure 4.3e. In Figure 4.3g, the selected endpoints for all directions are marked and in Figure 4.3h the result of the filtering and ellipse fitting are shown, which steps are discussed in Section 4.2.4.

4.2.4 Center point and diameter estimation

Earlier, it was assumed that we have a good pupil center point estimation. This estimation is produced by utilizing a geometrical relation between the center point and the chords of a circle. A random point is picked within the circle (shown in Figure 4.4a) and chords are drawn through this point (Figure 4.4b). The perpendicular bisectors of these chords intersect in one single position, which is the center of the circle (Figure 4.4c).

The estimation process starts by placing seed points on a grid. For each seed point having a low pixel intensity (likely pupil region), we do the followings. By utilizing the energy attenuation model based method, rays are being propagated in all directions from the currently investigated seed point. Then, the most probable boundary points are selected. Endpoints that correspond to rays with opposite directions are grouped to form chords. After filtering the chords based on their lengths, their perpendicular bisectors are calculated. Note that because of the inaccurate chord estimations, the bisector lines do not intersect in a single point. To handle this problem, a Least-Square solution is calculated to find the best possible intersection point for the given lines. Then, the distances between this newly estimated center point and the previously found boundary points are computed. The standard devia-

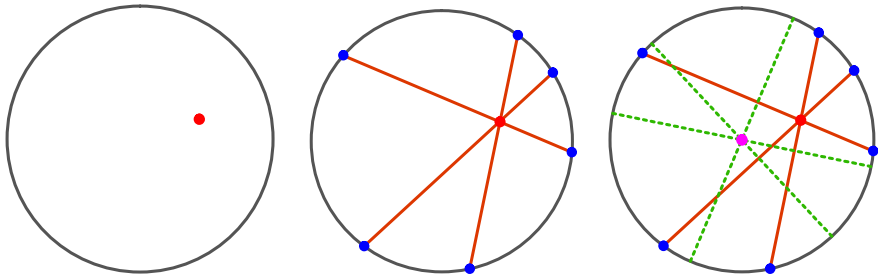


Figure 4.4: *Circle center point estimation: a) random point within a circle, b) chords through the selected point, c) perpendicular bisectors and the center point of circle.*

tion of these distances serves as a score for the corresponding estimated center point. After processing all the frames from the video, a large number of pairs of estimated positions and their scores are produced. The final output of the method is the center point having the minimal standard deviation score, i.e. it is the most accurate center point candidate.

The proposed estimation process is illustrated in Figure 4.5. The seed points marked by white circles were filtered out in the initial step because their pixel intensity was too high. The yellow plus sings mark chord endpoints that were considered as outliers based on median filtering and the green plus signs mark the remaining chords used during the estimation phase. It can be observed that from the selected seed point (marked by a black circle) the process estimated the center point (marked by a green star) with acceptable accuracy.

After the selection of the best estimated center point, each frame is processed by

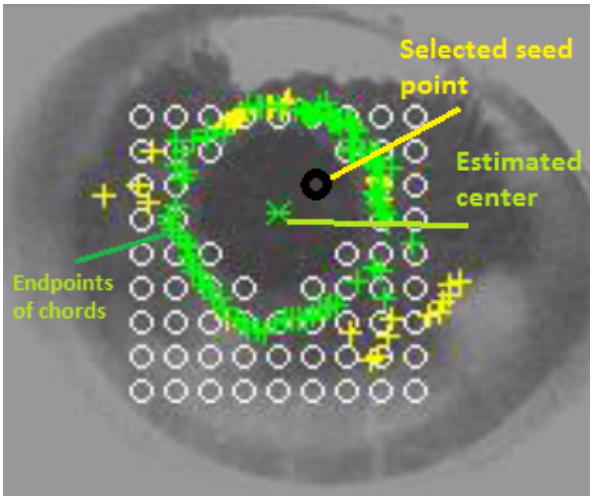


Figure 4.5: *Pupil center point estimation.*

the method described is Section 4.2.3 using this point as the source of radiation. The output for each frame is a set of endpoints and their distances from the center point, which may contain faulty detections due to high intensity reflections, undetectable boundary sections, and occlusions. These factors occur randomly compared to the regularity of the pupil boundary points. To utilize this difference, the endpoint distances are quantized and their histogram is calculated. From the histogram, the bin with the maximal magnitude is chosen, which is the most probable value of the pupil radius. Using this value, endpoint filtering is performed. Only endpoints with distances close to the previously chosen radius are kept. Due to the relative position of the animal to the camera, the pupil is usually distorted to an ellipse because of the non-perpendicular viewing angle. Based on the retained endpoints, Least-Squares-based ellipse fitting is performed. To produce a final measure of the pupil diameter,

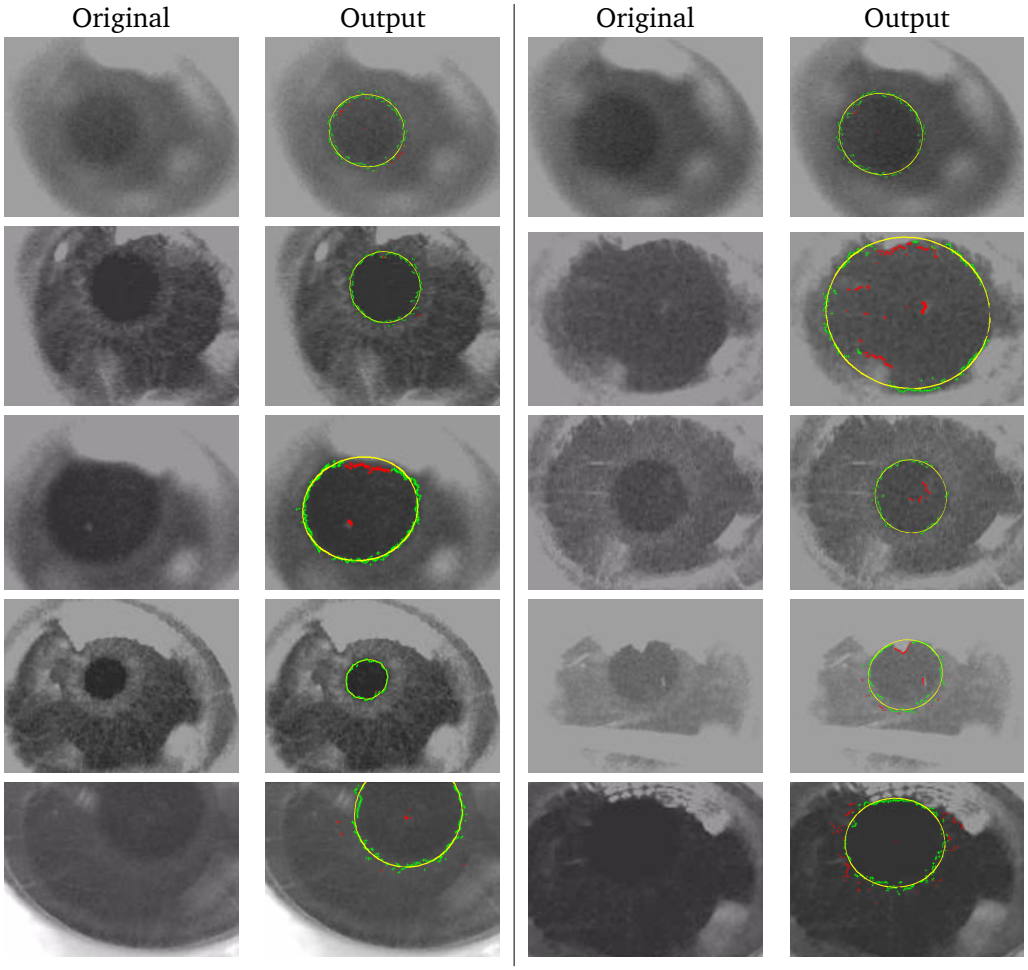


Figure 4.6: *Results of endpoint filtering and ellipse fitting. The green signs mark the retained endpoints, while the red ones the faulty endpoints. The yellow colored ellipses are fitted to the retained endpoints.*

the major axis length of the fitted ellipse is calculated. If there is not enough information (too few endpoints after filtering) to fit an ellipse, only the approximated radius length is used.

The output of the automated process is the pupillogram, which curve is produced by running the algorithm in each video frame. This raw pupillogram is post-processed with median filtering to eliminate outlier measurements. Sample result of the above-mentioned phases (filtering, radius measurement, and ellipse fitting phases) are shown in Figure 4.6. In each picture, the green marks represent the retained endpoints, while the red ones denote the faulty endpoints. The ellipses fitted to the retained endpoints are represented by yellow colored ellipses.

4.2.5 Evaluation of the pupil measurement method

The automated pupillometry was evaluated on 20 videos, each of them contained 450 frames. The outcome of the automated process was compared to values obtained by manual annotations and corrections. A graphical user interface was developed to support the correction task. After every fifth video frame, the software let the medical expert to correct the estimated pupil boundary by hand if it was necessary. It was reasonable to supervise only every fifth frame because in that case, the variance of intensity and pupil parameters was bigger compared to having the same number of supervised measurements in consecutive frames. It should be added that the pupil boundary was not always sharp enough to allow the placement of separating points exactly on it. Thus, acceptable measurement means that an expert would place the boundary point very close (within few pixels) to the automatically computed position. We compared our result only for the 1800 supervised frames and listed the results of statistical analysis in Table 4.1. The following metrics were computed. For each frame $f \in F_v$ in video $v \in V$ the relative percentage error (RPE) of the diameter measurement was defined as:

$$RPE_{vf} = 100\%\cdot\frac{|d_{vf} - \hat{d}_{vf}|}{|d_{vf}|},$$

where d_{vf} is the corrected diameter, and $\hat{d}_{vf}$ is the estimated pupil diameter; F_v is the set of indices of supervised frames in video $v \in V$, $V = \{1, 2, ..., 20\}$. In Table 4.1 the rows correspond to the aggregation of RPE for all frames per video, and the columns correspond to the aggregation of the per video performances for all videos in the test dataset. For example, for video v, the mean relative percentage error μRPE is calculated as

$$\mu RPE = \frac{1}{|F_v|}\sum_{f\in F_v} RPE_{vf},$$

Table 4.1: *Performance analysis of the proposed pupil measurement method. The rows correspond to the aggregation of RPE for all frames in a video, and the columns correspond to the aggregation of the per video performances for all videos in the test dataset.*

		Minimum	Maximum	Mean	Standard deviation	Median
All supervised frames	Maximum relative error (%)	0.48	22.41	10.64	5.32	8.79
	Mean relative error (%)	0.60	4.85	1.95	1.38	1.53
	Standard deviation of relative error (%)	0.95	6.79	2.68	1.35	2.44
	Median of relative error (%)	0.00	3.10	0.90	1.23	0.00
Corrected frames	Mean relative error (%)	1.41	8.52	4.13	1.63	4.13
	Standard deviation of relative error (%)	0.98	7.19	2.94	1.58	2.46
	Median of relative error (%)	1.02	7.32	3.53	1.63	3.30
	Corrected frames (%)	15.28	88.89	46.38	23.03	40.97

where $|F_v|$ is the number of supervised frames in video v. Then, these μRPE values were calculated for all videos, and for example, the maximum of them in our case was 4.85% (which appears in the 2nd data cell in the 2nd row of Table 4.1). This can be interpreted as the maximal value of the mean relative errors computed for all videos. The other values were calculated similarly but with different aggregation and selection methods. The upper part of Table 4.1 shows the result of the analysis for all supervised frames, while the bottom part shows the same metrics but for the corrected frames only. The last row gives information about the ratio of corrected frames.

As it can be observed, on the supervised frames the mean relative percentage error was less than 2%, and on the corrected frames, it was around 4%. This accuracy is sufficient to assist the medical staff in their work. In average, 40% of the frames were corrected by hand during the supervision phase, but in most cases there were only minor differences between the estimated and corrected diameter lengths. The proposed automatic method processed 3 frames per second, which included ray tracking, end-

point filtering, hierarchical clustering, ellipse fitting and output video rendering. An input video was processed in 3 minutes, which was a significant speed-up compared to manual annotation.

4.3 Utilization of the pupil measurements

The motivation behind the pupil measurements was clarified in the introduction of the chapter. In the current section, I highlight my contributions to the related medical results beyond the development of the automated pupillometry algorithm described in the previous sections. The medical results about schizophrenia-induced alterations in the PLR curves were published in a medical journal [10]. During these investigations two series of experiments were performed in sedated (n=54) and anesthetized (n=20) control Wistar (healthy), and test WISKET rats (with schizophrenia-like alterations). After a 10-minute long dark adaptation period, the recordings lasted for 15 s in sedated, and for 60 s in anesthetized animals. The animals were positioned close to the recording camera and an intensive visible light stimulus (approx. 300 cd/m^2 for 600 ms) was flashed into their left eyes. The IR-camera recorded pupillary responses at a speed of 24 frames-per-second under infrared illumination.

The videos were processed by the automated diameter measurement software that produced the corresponding pupillograms. To compare the responses of the healthy and WISKET animals, descriptive features from the PLR curves were extracted. These are relevant from the pathophysiological point of view and suitable to emphasize the differences between the groups regarding the autonomic nervous system activities. To support this feature extraction phase, an automated method was designed, which produced 40 features from a pupillogram. Many of these features were traditionally used parameters such as the initial diameter, which is the size of the pupil before the light impulse; minimum diameter; reaction time; maximum of the redilated diameter; etc. Several new features were also implemented to obtain information about the dynamics of the response, thus 11 velocity-related descriptors were introduced including the average, and the maximal contraction velocities and the times required to reach the latter. The instantaneous velocities of the redilation phase at different time points were also calculated.

Novel, smoothness related descriptors were introduced, as well. A polynomial curve with a given order was fitted to the redilation part of the response. The area between the original and fitted curves served as a measure of non-smoothness. Fifth order polynomials were chosen, which were flexible enough to follow the slow perturbations of the original curve and indicated only the short, abnormal swings. In Figure 4.7, a representative PLR response curve and a marked subset of the extracted features are presented.

These features served as the basis for data analysis that aimed to find differences

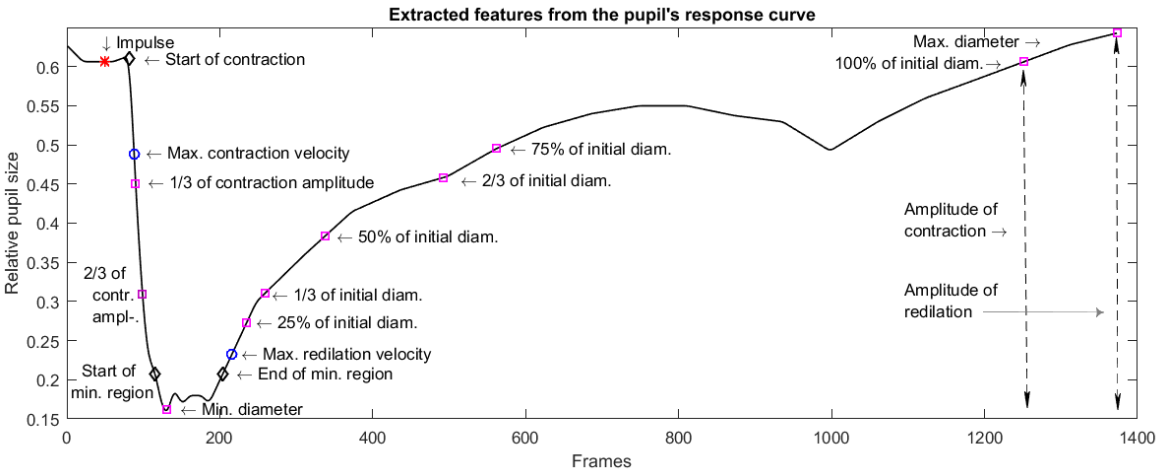


Figure 4.7: *Pupillary light reflex curve and a marked subset of the extracted features.*

between the control and test animals. In [10], the relationships between pupil parameters were assessed by linear regression analysis and by the calculation of the Pearson correlation coefficient. However, besides the investigation of the differences in PLR parameters, the other goal of the research was to use pupillometry as a quick examination to facilitate the selection of the animals during the breeding process. Therefore, I trained a decision tree to investigate the possibility of binary classification and the model was evaluated by using cross-validation.

The detailed discussion of the results of the statistical analysis is found in our recent study [10]. In accordance with these results, the trained decision tree selected almost the same features as predictor variables as the statistical analysis suggested.

In the sedated group, which had the greater cardinality, the fitted decision tree achieved 71% accuracy measured by 5-fold cross-validation. The algorithm selected the following predictors: minimum diameter, initial diameter, and average redilation speed (change of pupil size/frame time). With the combination of the two analysis methods, significant differences were noticed between the control and test groups. The initial and minimum pupil diameters were larger and the degree of the constriction was lower in the WISKET rats. The flatness of the curve (length of the minimum region) and the contraction time were shorter in the control group.

The analysis of the anesthetized animals showed that they cannot be divided into two classes reliably. It is assumed that while anesthesia can prevent stress and allows a convenient investigation of pupillary reactions for a longer period, it also diminishes the differences between the two groups in this autonomic response. The classifier achieved only 60% accuracy. The selected predictors were the amplitude of contraction, the average redilation speed, and the time required to reach the maximal redilation speed.

The results showed that anesthetized animals cannot be used in these experiments, so the attention was focused on the sedated animals. However, the sedated animals were still able to move their heads during the experiments, which affected negatively the recording of the pupil with the closely placed camera. Based on the decision tree analysis, the initial diameter and the minimum diameter seemed to be the most reliable features for the classification, although the dynamic, time-related features seemed to be less relevant (or more samples (rats) are required to detect the potential differences). From these observations, it could be concluded that a more robust measurement process was required, which had a relatively short period and provided multiple light stimuli to induce reoccurring reflex responses and minimum diameter occasions. thus enabling a complex and more detailed analysis.

4.4 Improved pupillometry

In the previous sections, a traditional pupillometry pipeline and its automation was explained, which led to significant medical results in the field of schizophrenia-related research. However, the experimentation setup set video quality and robustness limitations. In this section, a revised automated pupillometry is described that was improved by the rearrangement and hardware modification of the recording camera setup, and by the development of a new, data-driven AI-based pupil detection method.

4.4.1 Improved data acquisition

To improve the robustness of the recordings, the measurement setup was redesigned. An IR LED (infrared light-emitting diode) ring was attached around the camera lens, which configuration placed the camera optical axis and the illuminating IR LEDs (nearly) on the same axis. With this modification, the camera was able to capture the light reflected back from the retina. This reflected light produced the so-called ”Bright pupil effect” [61]. Here, the pupil region operates as a small ”window” that lets the incoming and reflected light through, thus it glows bright, while the other structures scatter the incoming light and remain dark in the recorded image. This effect enhanced the signal-to-noise ratio and improved the detectability of the pupil region. This increase allowed the camera to be placed farther from the eye, thus tiny animal movements did not affect the recording quality. In Figure 4.8 the differences between the old and rearranged setups are presented, while Figure 4.8b illustrates the improved image quality.

A new embedded system was also developed to schedule the experiments. This hardware can be controlled from a graphical software running on the PC to initiate different visible light impulse sequences. The scheduler implements accurate time

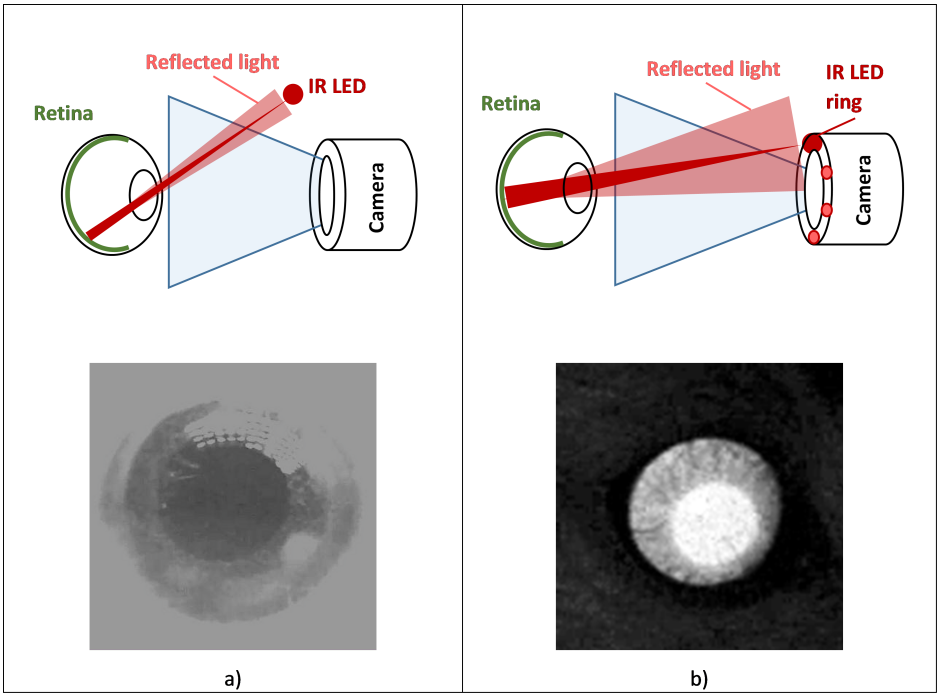


Figure 4.8: *Difference between the old (a) and the rearranged (b) measurement setups. a) The camera is close to the eye and the illuminating infrared LED placed far from the camera's optical axis. The camera records black pupil region (reflected light never reaches the camera). b) The camera is placed farther from the eye and the illuminating LEDs are close to the optical axis. The camera records bright pupil region (reflected light reaches the camera).*

synchronization and manages the timing of light stimuli. An RGB LED was integrated, which allows us to produce different light intensities and colors. This system makes future experiments flexible and complex relations can be investigated with more possible options.

4.4.2 Pupil segmentation dataset

The images recorded with the new measurement setup had a completely different nature (different intensity levels, different resolution, etc.), as can be seen in Figure 4.8b. Therefore, our previously detailed method could not be used. To support the development of a new pupil segmentation algorithm and to validate the new experimentation setup, 56 videos were recorded each containing more than 5000 frames. From these videos, 2564 randomly chosen frames were manually annotated (ellipses were manually fitted to the pupil regions) and collected into a training dataset. An additional set of 329 images were selected and annotated to form a challenging test dataset.

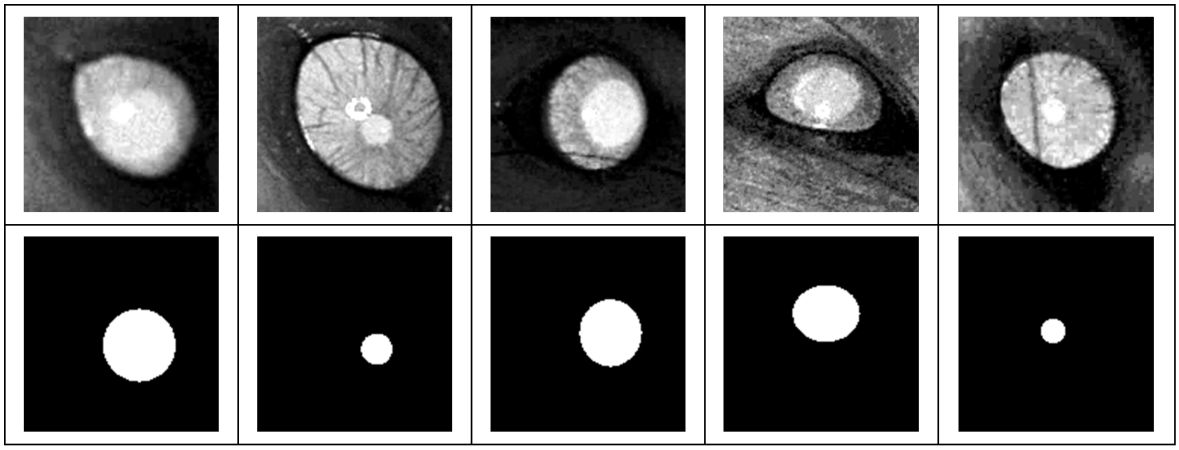


Figure 4.9: *Example image samples from the pupil segmentation dataset. Top row: original samples. Bottom row: corresponding binary pupil mask images.*

Traditional image processing methods were implemented to detect the eye region in the image, to crop this area, and to enhance the contrast. The resulting dataset contains 128×128 pixels-sized preprocessed images and the corresponding binary pupil masks. Figure 4.9 presents some samples and the corresponding pupil masks from the pupil segmentation dataset. The dataset is publicly available [67].

4.4.3 Data-driven pupil segmentation

The improved video quality reduced the complexity of pupil segmentation, thus, a reasonable amount of data was enough to train a data-driven method. Neural networks offer a modern way to solve a segmentation problem. Fully-convolutional networks are frequently used in similar problems as was explained earlier in this work. For the pupil segmentation task, a U-shaped structure, called the U-net [71] designed for biomedical applications was utilized. In a nutshell, this structure compresses and converts the input image to smaller feature maps and then tries to reconstruct the desired output, in our case, the binary mask. This reconstruction from the compressed representations happens through consecutive up-sampling steps. These down- and up-sampling layers form a symmetric structure that resembles to a capital letter *U* (this is where the name comes from). The key idea is to use layers, called skip layers, to connect the corresponding points of the U-shaped structure, which provide the compression-side, high-resolution feature maps to the up-sampling kernels. Without the details, a quick look at Figure 4.10 may help the reader to imagine the concept.

The training dataset contained input images with 128×128 resolution. The originally published version on the U-net worked with different input sizes (512×512), thus adaptations in the input layer and in the corresponding layers were required.

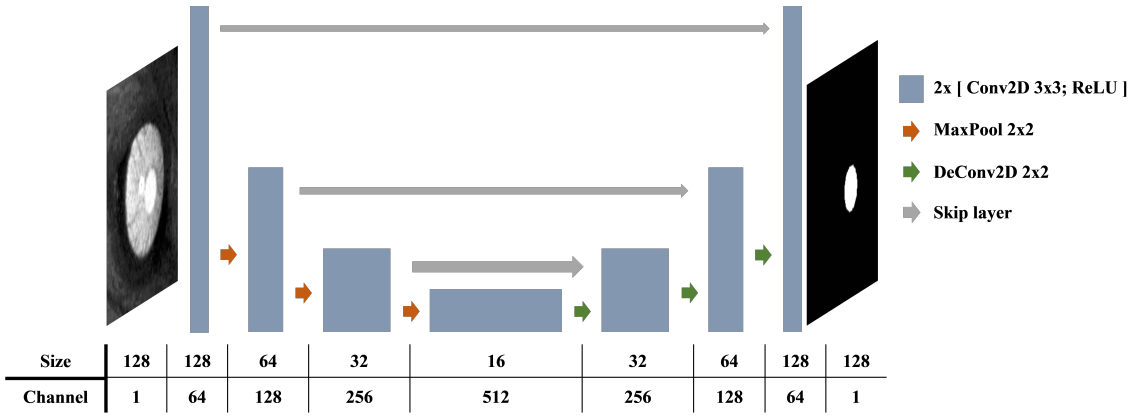


Figure 4.10: *The U-net neural network structure similar to [71].*

In the down-sampling part, three convolutional and max-pooling layers were applied and the number of channels was doubled after each pooling, as the original paper suggested. At the bottom of the U-shaped structure, 512 pieces of 16×16 sized feature maps were calculated. The abstract visualization of the used structure can be observed in Figure 4.10. Instead of the random initialization of the ”de-convolutional” layers, these weights were initialized to perform bilinear up-sampling. The predicted and ground-truth pupil masks were compared by binary cross-entropy (the loss function). Adam method [48] was used as the optimizer and the batch size was 64. Data augmentation was employed to prevent model overfitting. These methods included random flips, and random crop and resize. No other regularization techniques were used, which is a common direction in fully-convolutional networks. The training process was optimized based on the analysis of the validation loss curves, where 10% of the training dataset served as the validation set. Early-Stopping was applied to terminate the training process when the validation loss stopped decreasing. This setup was tested with different learning rates. The best performing model ran for 200 epochs, with the learning rate of 0.001. The algorithm was implemented in PyTorch [63].

The output of the neural network was an almost binary image. The pixel values were close to 1.0 when a pixel was considered as part of the pupil. The output was binarized with an empirically set threshold value — 0.97 during these experiments. In the resulting binary images, the object contours were extracted and ellipses were fitted to these boundary points. The final output was the filled, solid mask of the fitted ellipse. If multiple ellipses were identified, the false detections were filtered out by simple rules based on the sizes, shapes, and locations of the ellipses. A set of sample images and the predicted ellipses (without false detection filtering) visualized by red contours are presented in Figure 4.11.

The trained model was evaluated on the test dataset containing 329 images. The

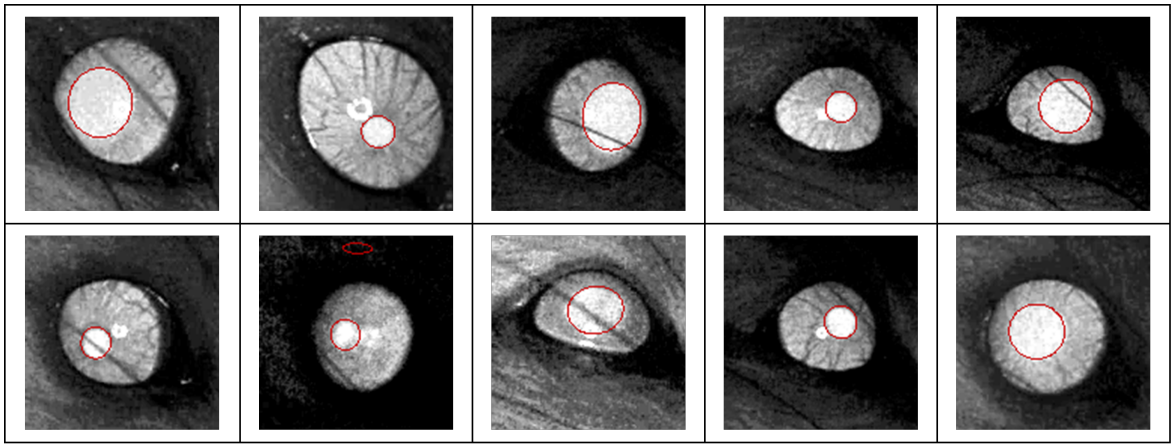


Figure 4.11: *Example result images. Red ellipses present the predicted ellipses' borders.*

Jaccard-index (Intersection over Union) was computed to compare the ground-truth and predicted pupil masks. The average Jaccard-index value on the test dataset was 0.815. The test set is challenging and there were samples, which were completely misclassified by the algorithm. Therefore, the median value was also calculated, which measure is more robust to the outliers. The median relative Jaccard-index was 0.883.

Besides the pupil mask comparison, the major axis lengths (diameters) of the original and predicted ellipses were also compared, which was important, because the desired output of the pupillometry was the pupillogram, the curve of pupil diameters. The average relative pupil diameter error was 12%, the median relative error was 4%. In Figure 4.12a the sorted original and the corresponding predicted diameters are presented. It can be seen that most of the error occurred when the pupil diameter was smaller than 20 pixels. This information loss might be caused by the down-sampling section of the U-shaped structure. The resulting error characteristics show a structure that might be utilized later during post-processing to reduce the inaccuracy at the smaller pupil diameters. In Figure 4.12b the relationship between the Jaccard-index and the relative diameter error can be observed. Small Jaccard-index value not necessarily implies considerable relative pupil diameter estimation error. This can be explained because if only a small portion of the mask is missing, this will reduce the Jaccard-index value (as it is the error of the mask). However, it is possible that the missing part only has an influence on the minor axis length of the ellipse, and the major axis length – the extracted diameter value – remains accurate.

The diameter-measurement accuracy of the AI-based solution is comparable to the traditional method described in the first part of the chapter. Both methods, however, could be improved with the utilization of post-processing, because additional temporal filtering is available if the algorithms are being run on consecutive

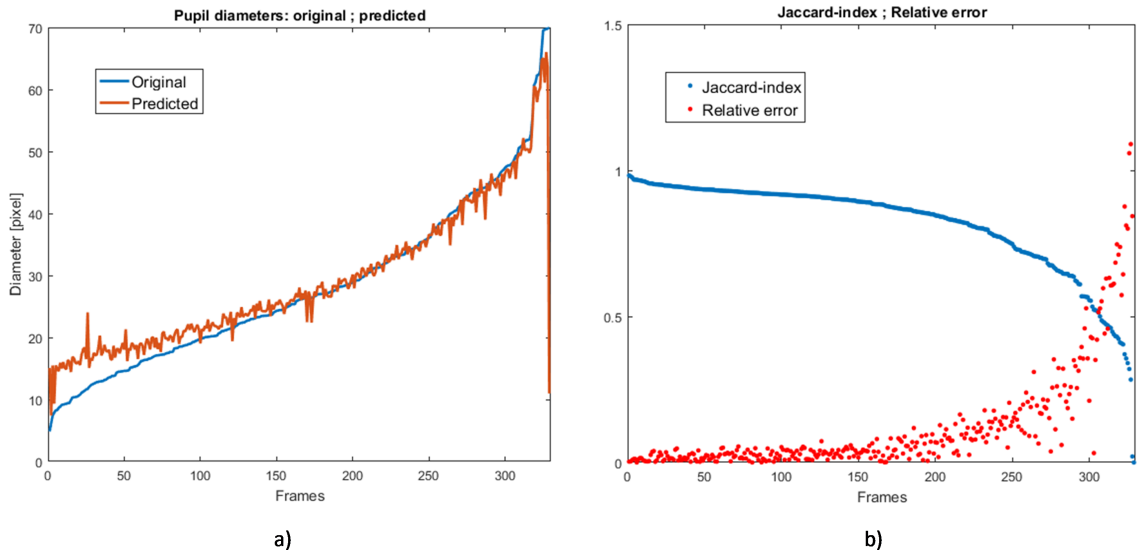


Figure 4.12: *a) Sorted original diameters and the corresponding predicted diameters. b) Sorted Jaccard-indecies and the corresponding relative diameter errors.*

video frames. This time-domain correlation could be utilized in more complex neural network structures, too, which would take a part of a video as the input (3D format). This possibility will be investigated in future work. Also, the new measurement method runs on a GPU-enabled PC, which provides a significant speed boost compared to the traditional algorithm and real-time processing can be achieved.

Besides the image quality improvements, the pupillometry with the new experimentation setup has a positive impact on the classification of rats, too. As the traditional examination procedure (one light impulse/experiment) is inherited in the extended measurements (light impulse sequences), at least the same set of features can be extracted from the pupillograms. However, the relations between the consecutive impulse responses may unfold more complex details about the alterations of the autonomic nervous system.

4.5 Summary and conclusions

In this chapter, the automation process of a pupillometry application was presented. The related medical research aims to reveal objectively detectable effects of schizophrenia on the autonomic nervous system through the examination of the pupillary light reflex in a rat model.

Traditional pupillometry experiments recorded the pupil reflex to light stimuli. Then, the pupil diameter was measured in each video frame to produce the pupillogram. To support the medical research and speed-up the tiring work of manual

pupil region annotation, an automated process to measure the pupil diameter was developed. To accurately detect the contour of the pupil, a novel image processing method, an energy attenuation based ray propagation algorithm was introduced. It handles the wide spectra of intensity parameters, the low contrast difference, the blur, and the occlusions, thus low-quality videos could be processed. The proposed method was evaluated on 20 videos and achieved an overall relative pupil diameter error around 2%.

With the utilization of the developed method, and based on the produced pupillograms, significant medical results were achieved that revealed altered autonomic nervous system functionalities in the investigated rat model. Nevertheless, the experimentation setup was sensitive to the rat motions that prevented longer and more complex examinations. The video quality was also poor. To overcome these limitations, the measurement process was redesigned to induce the bright pupil effect, which offered more robustness to the experiments. A hardware extension reduced the motion-sensitivity and increased the signal-to-noise ratio making the pupil segmentation problem simpler. However, because of the modified image properties, a new pupil detection algorithm was required. To develop such a data-driven segmentation algorithm, a publicly available pupil segmentation dataset was created. A fully-convolutional neural network was trained to produce binary pupil masks. On the test dataset, the relative pupil diameter predictor achieved 4% median error. The redesigned camera setup and the corresponding measurement software offer more robust animal experimentation and a fast, accurate data analysis.

4.6 Contributions

The author of this PhD thesis is responsible for the following contributions:

III/1. I developed and evaluated a pupil measurement method that is based on an energy attenuation model, implemented an automated feature extractor, and introduced new pupillogram features.

III/2. I redesigned the previously used pupillometry experimentation setup with a hardware extension on the camera, which enhanced the video quality, and thus supports more robust and more efficient experimentation.

III/3. I trained a fully-convolutional neural network for pupil segmentation that efficiently processes the videos acquired by the revised experimentation setup.

Bibliography

[1] Talal Ahmed, Momin Uppal, and Abubakr Muhammad. Improving efficiency and reliability of gunshot detection systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 513–517. IEEE, 2013.

[2] Shahin Amiriparian, N Cummins, S Julka, and BW Schuller. Deep convolutional recurrent neural network for rare acoustic event detection. In *Proc. DAGA*, pages 1522–1525, 2018.

[3] Victor Ayala-Ramirez, Carlos H. Garcia-Capulin, Arturo Perez-Garcia, and Raul E. Sanchez-Yanez. Circle detection on images using genetic algorithms. *Pattern Recognition Letters*, 27(6):652–657, 2006.

[4] Karl-Jürgen Bär, Michael Karl Boettger, Steffen Schulz, Christina Harzendorf, Marcus Willy Agelink, Vikram K Yeragani, Prtap Chokka, and Andreas Voss. The interaction between pupil function and cardiovascular regulation in patients with acute schizophrenia. *Clinical Neurophysiology*, 119(10):2209–2213, 2008.

[5] Karl-Jürgen Bär, Mandy Koschke, Michael Karl Boettger, Sandy Berger, Alexander Kabisch, Heinrich Sauer, Andreas Voss, and Vikram K Yeragani. Acute psychosis leads to increased qt variability in patients suffering from schizophrenia. *Schizophrenia research*, 95(1):115–123, 2007.

[6] Leo Leroy Beranek and Tim Mellow. *Acoustics: sound fields and transducers*. Academic Press, 2012.

[7] Jona Beysens, Alessandro Chiumento, Min Li, and Sofie Pollin. Touchspeaker, a multi-sensor context-aware application for mobile devices: from application to implementation. *Journal of Signal Processing Systems*, 90(10):1469–1478, 2018.

[8] Jona Beysens, Alessandro Chiumento, Sofie Pollin, and Min Li. Touchspeaker, a multi-sensor context-aware application for mobile devices. In *2016 IEEE international workshop on signal processing systems (SiPS)*, pages 23–26. IEEE, 2016.

[9] John Borwick. *Loudspeaker and headphone handbook*. CRC Press, 2012.

[10] Alexandra Büki, György Kalmár, Gabriella Kékesi, László G Nyúl, and Gyöngyi Horváth. Impaired pupillary control in "schizophrenia-like" WISKET rats. *Autonomic Neuroscience: Basic and Clinical*, 2018. Under revision.

[11] Bradley J. Cardinale et al. Biodiversity loss and its impact on humanity. *Nature*, 486(7401):59, 2012.

[12] Teh-Chuan Chen and Kuo-Liang Chung. An efficient randomized algorithm for detecting circles. *Computer Vision and Image Understanding*, 83(2):172–191, 2001.

[13] Yan Chen, Qian Guo, Xinyan Liang, Jiang Wang, and Yuhua Qian. Environmental sound classification with dilated convolutions. *Applied Acoustics*, 148:123 – 132, 2019.

[14] Yukun Chen, Yichi Zhang, and Zhiyao Duan. Dcase2017 sound event detection using convolutional neural network. *Detection and Classification of Acoustic Scenes and Events*, 2017.

[15] J.M. Cho, S.J. Lee, J.K. Kim, H.H. Choi, O.S. Kwon, and J.W. Kwon. A pupil center detection algorithms for partially-covered eye image. *2004 IEEE Region 10 Conference TENCON 2004.*, A:183–186, 2004.

[16] Francois Chollet et al. Keras. `https://keras.io`, 2015.

[17] Erik Cuevas, Noé Ortega-Sánchez, Daniel Zaldivar, and Marco Pérez-Cisneros. Circle detection by harmony search optimization. *Journal of Intelligent & Robotic Systems*, 66(3):359–376, 2012.

[18] Erik Cuevas, Valentín Osuna-Enciso, Fernando Wario, Daniel Zaldívar, and Marco Pérez-Cisneros. Automatic multiple circle detection based on artificial immune systems. *Expert Systems with Applications*, 39(1):713–722, 2012.

[19] Erik Cuevas, Felipe Sención-Echauri, Daniel Zaldivar, and Marco Pérez-Cisneros. Multi-circle detection on images using artificial bee colony (abc) optimization. *Soft Computing*, 16(2):281–296, 2012.

[20] Erik Cuevas, Daniel Zaldivar, Marco Pérez-Cisneros, and Marte Ramírez-Ortegón. Circle detection using discrete differential evolution optimization. *Pattern Analysis and Applications*, 14(1):93–107, 2011.

[21] Sambarta Dasgupta, Swagatam Das, Arijit Biswas, and Ajith Abraham. Automatic circle detection on digital images with an adaptive bacterial foraging algorithm. *Soft Computing*, 14(11):1151–1164, 2010.

[22] Dayton Audio DMA58-4, full-range loudspeaker. http://www.loudspeakerdatabase.com/Dayton/DMA58#4%CE%A9 [Online; accessed 2020-02-03.].

[23] Dayton Audio DMA105-8, full-range loudspeaker. http://www.loudspeakerdatabase.com/Dayton/DMA105#8%CE%A9 [Online; accessed 2020-02-03].

[24] Dayton Audio E220CF-8, woofer loudspeaker. "http://www.loudspeakerdatabase.com/Dayton/E220CF-8". [Online; accessed 2020-02-03].

[25] Douglas Coimbra de Andrade, Sabato Leo, Martin Loesener Da Silva Viana, and Christoph Bernkopf. A neural attention model for speech command recognition. *arXiv preprint arXiv:1808.08929*, 2018.

[26] A De Santis and D Iacoviello. Optimal segmentation of pupillometric images for estimating pupil shape parameters. *Computer methods and programs in biomedicine*, 84(2-3):174–87, 2006.

[27] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *arXiv preprint arXiv:1808.05377*, 2018.

[28] James A. Estes et al. Trophic downgrading of planet Earth. *Science*, 333(6040):301–306, 2011.

[29] F Alton Everest. *Master handbook of acoustics*. ASA, 2001.

[30] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters*, 65:22 – 28, 2015.

[31] Chollet Francois. *Deep learning with Python*. Manning Publications Co., 2018.

[32] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection in noisy environments. In *2007 15th European Signal Processing Conference*, pages 1216–1220, Sept 2007.

[33] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[34] Gore Portable Electronic Acoustic Vents. https://www.gore.com/sites/g/files/ypyipe116/files/2016-12/GORE_Acoustic_Vents_immersion_EN_DEC_2016.pdf.
Accessed: 2019-12-03.

[35] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354 – 377, 2018.

[36] Mordechai Guri, Yosef Solewicz, Andrey Daidakulov, and Yuval Elovici. SPEAKE(a)R: Turn Speakers to Microphones for Fun and Profit. In *11th USENIX Workshop on Offensive Technologies (WOOT 17)*, Vancouver, BC, 2017. USENIX Association.

[37] Gyorgy Kalmar. Source code for the paper: Analysis and Utilization of Reverse Mode Loudspeakers. `https://github.com/Gyoorey/reverse_mode_speakers`.

[38] Haggai Hermesh, Roni Shiloh, Yoram Epstein, Hillel Manaim, Abraham Weizman, and Hanan Munitz. Heat intolerance in patients with chronic schizophrenia maintained with antipsychotic drugs. *American Journal of Psychiatry*, 157(8):1327–1329, 2000.

[39] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[40] Gyöngyi Horváth, Péter Liszli, Gabriella Kékesi, Alexandra Büki, and György Benedek. Characterization of exploratory activity and learning ability of healthy and 'schizophrenia-like' rats in a square corridor system (ambitus). *Physiology & behavior*, 169:155–164, 2017.

[41] Gyöngyi Horváth, Zita Petrovszki, Gabriella Kékesi, Gábor Tuboly, Balázs Bodosi, János Horváth, Péter Gombkötő, György Benedek, and Attila Nagy. Electrophysiological alterations in a complex rat model of schizophrenia. *Behavioural brain research*, 307:65–72, 2016.

[42] Frederick Vinton Hunt. *Electroacoustics: The Analysis of Transduction, and Its Historical Background (Harvard Monographs in Applied Science)*. Cambridge, MA, USA: Harvard University press, 1954.

[43] J. Illingworth and J. Kittler. A survey of the hough transform. *Computer Vision, Graphics, and Image Processing*, 44(1):87–116, 1988.

[44] Iridium Short Burst Data (SBD). `https://www.iridium.com/services/iridium-sbd/`. Accessed: 2019-12-11.

[45] Chieh-Chi Kao, Weiran Wang, Ming Sun, and Chao Wang. R-crnn: Region-based convolutional recurrent neural network for audio event detection. *arXiv preprint arXiv:1808.06627*, 2018.

[46] Gabriella Kékesi, Zita Petrovszki, György Benedek, and Gyöngyi Horváth. Sex-specific alterations in behavioral and cognitive functions in a "three hit" animal model of schizophrenia. *Behavioural brain research*, 284:85–93, 2015.

[47] Jieun Kim and Kyungmo Park. An image processing method for improved pupil size estimation accuracy. *Engineering in Medicine and Biology Society, . . .* , 1:720–723, 2003.

[48] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[49] Lawrence E. Kinsler, Austin R. Frey, Alan B. Coppens, and James V. Sanders. *Fundamentals of Acoustics*. Wiley, 1999.

[50] Eunchong Lee, Hyunsoo Kim, and Ji Won Yoon. Various threat models to circumvent air-gapped systems for preventing network attack. In *Revised Selected Papers of the 16th International Workshop on Information Security Applications - Volume 9503*, WISA 2015, page 187–199, Berlin, Heidelberg, 2015.

[51] J C Lee, J E Kim, and K M Park. Evaluation of the methods for pupil size estimation : On the perspective of Autonomic Activity. *Proceedings of the 26th Annual International Conference of he IEEE EMBS*, pages 1501–1504, 2004.

[52] YD Li, ZB Hao, and Hang Lei. Survey of convolutional neural network. *Journal of Computer Applications*, 36(9):2508–2515, 2016.

[53] Arnold Lidsky, Gad Hakerem, and Samuel Sutton. Pupillary reactions to single light pulses in psychiatric patients and normals. *Journal of Nervous and Mental Disease*, 1971.

[54] Hyungui Lim, Jeongsoo Park, and Yoonchang Han. Rare sound event detection using 1d convolutional recurrent neural networks. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, pages 80–84, 2017.

[55] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11 – 26, 2017.

[56] C. Lopez-Tello and V. Muthukumar. Classifying acoustic signals for wildlife monitoring and poacher detection on UAVs. In *2018 21st Euromicro Conference on Digital System Design (DSD)*, pages 685–690, Aug 2018.

[57] S. Srinivasan M. R. Azimi-Sadjadi, Y. Jiang. Acoustic classification of battlefield transient events using wavelet sub-band features. *Proc.SPIE*, 6562:6562 – 6562 – 8, 2007.

[58] Justin Salamon Michael Mandel and Daniel P.W. Ellis. *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. New York University, 2019.

[59] D.B. Miller, B.J. Benton, S.W. Hulet, R.J. Mioduszewski, C.E. Whalley, J.C. Carpin, and S.A. Thomson. An image analysis method for quantifying elliptical and partially obstructed pupil areas in response to chemical agent vapor exposure. *2003 IEEE 29th Annual Proceedings of Bioengineering Conference*, pages 63–64, 2003.

[60] Kabhilan Mohan, Matthew M Harper, Helga Kecova, Eun-Ah Ye, Tatjana Lazic, Donald S Sakaguchi, Randy H Kardon, and Sinisa D Grozdanic. Characterization of structure and function of the mouse retina using pattern electroretinography, pupil light reflex, and optical coherence tomography. *Veterinary ophthalmology*, 15(s2):94–104, 2012.

[61] C.H Morimoto, D Koons, A Amir, and M Flickner. Pupil detection and tracking using multiple light sources. *Image and Vision Computing*, 18(4):331 – 335, 2000.

[62] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, April 2015.

[63] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[64] Zita Petrovszki, Gábor Ádám, Gábor Tuboly, Gabriella Kékesi, György Benedek, Szabolcs Keri, and Gyöngyi Horváth. Characterization of gene–environment interactions by behavioral profiling of selectively bred rats: The effect of NMDA receptor inhibition and social isolation. *Behavioural brain research*, 240:134–145, 2013.

[65] Karol J Piczak. Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2015.

[66] Mark D Plumbley, Christian Kroos, Juan P Bello, Gaël Richard, Daniel PW Ellis, Annamaria Mesaros, et al. *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*. Tampere University of Technology, 2018.

[67] Pupil segmentation dataset. `www.github.com/Gyoorey/PupilDataset`.

[68] QinetiQ ears gunshot localization system datasheet. `https://qinetiq-na.com/wp-content/uploads/Datasheet_SWATS_LR.pdf`.

[69] Ali Ajdari Rad, Karim Faez, and Navid Qaragozlou. Fast circle detection using gradient pair vectors. In *DICTA*, pages 879–888, 2003.

[70] Nicola Ritter, James Cooper, Robyn Owens, and Paul P. Van Saarloos. Location of the pupil-iris border in slit-lamp images of the cornea. *Proceedings - International Conference on Image Analysis and Processing, ICIAP 1999*, pages 740–745, 1999.

[71] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015.

[72] Emily S Rothwell, Fred B Bercovitch, Jeff RM Andrews, and Matthew J Anderson. Estimating daily walking distance of captive african elephants using an accelerometer. *Zoo biology*, 30(5):579–591, 2011.

[73] Brian M Sadler, Tien Pham, and Laurel C Sadler. Optimal and wavelet-based shock wave detection and estimation. *The Journal of the Acoustical Society of America*, 104(2):955–963, 1998.

[74] J. Salamon and J. P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, March 2017.

[75] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA, Nov. 2014.

[76] Christie Sampson, John McEvoy, Zaw Min Oo, Aung Myo Chit, Aung Nyein Chan, David Tonkyn, Paing Soe, Melissa Songer, A. Christy Williams, Klaus Reisinger, George Wittemyer, and Peter Leimgruber. New elephant crisis in Asia—Early warning signs from Myanmar. *PLOS ONE*, 13(3):1–13, 03 2018.

[77] Savannah Tracking website. http://www.savannahtracking.com. Accessed: 2019-12-03.

[78] ShotSpotter website. www.shotspotter.com/products/military.html.

[79] Gyula Simon, Miklós Maróti, Ákos Lédeczi, György Balogh, Branislav Kusý, András Nádas, Gábor Pap, János Sallai, and Ken Frampton. Sensor network-based countersniper system. In *In Proc. of ACM SenSys*, pages 1–12, New York, NY, USA, 2004. ACM Press.

[80] Richard H Small. Direct radiator loudspeaker system analysis. *Journal of the Audio Engineering Society*, 20(5):383–395, 1972.

[81] Joseph Soltis, Rory P Wilson, Iain Douglas-Hamilton, Fritz Vollrath, Lucy E King, and Anne Savage. Accelerometers in collars identify behavioral states in captive african elephants loxodonta africana. *Endangered Species Research*, 18(3):255–263, 2012.

[82] STM32Cube initialization code generator. https://www.st.com/content/st_com/en/products/development-tools/software-development-tools/stm32-software-development-tools/stm32-configurators-and-code-generators/stm32cubemx.html. Accessed: 2019-12-03.

[83] STM32L476RG ultra-low-power microcontroller. https://www.st.com/en/microcontrollers/stm32l476rg.html. Accessed: 2019-12-03.

[84] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Learning from between-class examples for deep sound recognition. *arXiv preprint arXiv:1711.10282*, 2017.

[85] Carlo Tomasi and Takeo Kanade. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[86] Stanley M Tomkiewicz, Mark R Fuller, John G Kie, and Kirk K Bates. Global positioning system and associated technologies in animal behaviour and ecological research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1550):2163–2176, 2010.

[87] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 21–26, Sep. 2007.

[88] Peter Volgyesi, Gyorgy Balogh, Andras Nadas, Christopher Nash, and Akos Ledeczi. Shooter localization and weapon classification with soldier-wearable networked sensors. *5th International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2007.

[89] Wake-on-sound microphone. `http://www.puiaudio.com/product-detail.aspx?categoryId=4&partnumber=PMM-3738-VM1010-R`. Accessed: 2019-12-03.

[90] Jake Wall, George Wittemyer, Brian Klinkenberg, and Iain Douglas-Hamilton. Novel opportunities for wildlife conservation and research with real-time monitoring. *Ecological Applications*, 24(4):593–601, 2014.

[91] G.B. Whitham. Flow pattern of a supersonic projectile. *Communications on pure and applied mathematics*, 5(3):301, 1952.

[92] Christopher C. Wilmers, Barry Nickel, Caleb M. Bryce, Justine A. Smith, Rachel E. Wheat, and Veronica Yovovich. The golden age of bio-logging: how animal-borne sensors are advancing the frontiers of ecology. *Ecology*, 96(7):1741–1753, 2015.

[93] George Wittemyer, Guillaume Bastille-Rousseau, and Joseph M. Northrup. Behavioral valuation of landscapes using movement data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2019. In Press.

[94] Peter H. Wrege, Elizabeth D. Rowland, Sara Keen, and Yu Shiu. Acoustic monitoring for conservation in tropical forests: examples from forest elephants. *Methods in Ecology and Evolution*, 8(10):1292–1301, 2017.

[95] HK Yuen, J Princen, J Illingworth, and J Kittler. Comparative study of hough transform methods for circle finding. *Image and Vision Computing*, 8(1):71–77, 1990.

[96] Theodore P Zahn and David Pickar. Autonomic activity in relation to symptom ratings and reaction time in unmedicated patients with schizophrenia. *Schizophrenia research*, 79(2):257–270, 2005.

[97] Danjie Zhu, Steven T. Moore, and Theodore Raphan. Robust pupil center detection using a curvature algorithm. *Computer Methods and Programs in Biomedicine*, 59(3):145–157, 1999.

Summary

The PhD thesis presents three data analysis applications including embedded audio classification systems and image segmentation methods. A common approach connects them that is with hardware-oriented modifications and software co-development the high-level tasks became feasible, easier, or more accurate.

The dissertation consists of three major parts. In Chapter 2, an animal-borne gunshot detector system is presented that was improved by a novel wake-up mechanism. Another acoustic event detection related topic is investigated in Chapter 3 that studies the loudspeakers' sound recording capabilities, which option became feasible with a simple hardware extension. In Chapter 4, a video processing application is detailed that automated and improved the pupillometry of rats to support a schizophrenia-related medical research in a rat model.

Animal-Borne Anti-Poaching System

Poaching is listed among the top five drivers of biodiversity loss. Interventions to reduce it typically follow classic law enforcement approaches, however, poaching of the wildlife tends to occur in remote areas with low human densities, where detection is difficult. In addition, poaching of large, high-value species is militarized and supported by global crime syndicates. As such, local wildlife agents are operationally overwhelmed, not only in terms of law enforcement equipment, but often due to the limited capacity to monitor widely distributed animals. The development of technologies designed to overcome the challenges of remote wildlife protection is needed. Nowadays, a promising direction is the utilization of animal-borne sensors, particularly GPS-equipped collars, which are used to enhance real-time wildlife protection.

In Chapter 2, an animal-borne acoustic gunshot detector was introduced that extended the functionality of widely-used GPS tracking collars. With the fusion of the two systems, gunshot alerts can be raised in real-time coupled with location data. The main challenges were the multi-year lifetime requirement, the preservation of the recorded ballistic shockwave sound quality, and the minimization of the false positive gunshot detection rate. With a novel acoustic delay line structure that uti-

lizes two microphones with different characteristics, the power consumption reduced by 88% and the detection accuracy improved significantly. The hardware and software of the gunshot detector module were optimized to fit into GPS tracking collars used for elephants, and the first prototype was deployed on a wild elephant. To evaluate the system, controlled real-world experiments were also carried out. A dataset was collected containing environmental sounds, mechanical impact noises, and real gunshots. The proposed gunshot detection algorithm achieved good results on this dataset as all the gunshots were successfully detected and no false positive alarms were generated. Data-driven methods were also briefly mentioned and a randomized architecture-search algorithm was developed that generated, trained, and compared 1D and 2D convolutional neural networks for gunshot detection.

Reverse Mode Speakers

Chapter 3 investigated the ”microphone” mode of loudspeakers (referred to as reverse mode). In this state, the speaker converts sound to electrical signals, thus its environment becomes observable. The proposed idea was the utilization of reverse mode speakers in acoustic event detection applications. The hardware extension that offers this extra functionality is minimal and can be implemented by a simple embedded device. This device provides the original, radiating mode operation but extra, microphone-like capabilities also become feasible. For example, such extended speakers could be employed in security applications, where suspicious acoustic event detection is required.

The work introduced the analysis of the reverse mode through the formation of an equivalent mechanical circuit and included the results of experiments, simulations, and measurements. Possible utilization perspectives of the reverse mode speakers were also proposed in the chapter, and I designed and implemented an audio event detector device based on the reverse mode, and demonstrated its use by a simple, data-driven clap detector. These investigations assumed that during the sound-capturing phase, the loudspeaker was idle, meaning that its driving source was inactive. However, the work also included the analysis of the more challenging active reverse mode, when the speakers may be used for acoustic event detection while they are actively radiating sound.

Automated Pupillometry

In Chapter 4, the automation of a pupillometry application was presented. The related medical research aims to reveal objectively detectable effects of schizophrenia on the autonomic nervous system through the examination of the pupillary light reflex in a rat model.

Traditional pupillometry experiments record the pupil reflex to light stimuli and then, the pupil diameter is measured in each video frame to produce the pupillogram. To support the medical research and to speed-up the work of pupil region annotation, an automated image processing method was developed to measure the pupil diameter. A novel, energy attenuation model based ray propagation algorithm was introduced to accurately detect the contour of the pupil. It can handle the wide spectra of intensity parameters, the low contrast, the motion-blur, and the occlusions, thus low-quality videos could be processed. We evaluated the proposed method on 20 videos, which achieved an overall relative pupil diameter error around 2%.

With the proposed automated method, significant medical results were presented that revealed altered autonomic nervous system functionality and impaired pupillary control in the investigated rat model. To overcome various technical limitations, the experimentation setup was redesigned, which then offered more robustness to the animal experiments. A hardware extension on the camera reduced the motion-sensitivity and increased the signal-to-noise ratio that led to high-quality videos, thus made the pupil segmentation problem simpler. This problem was solved by a fully-convolutional neural network, which was trained on our publicly available pupil segmentation dataset. On the test images, the pupil diameter predictor achieved 96% accuracy.

Contributions of the thesis

In the **first thesis group**, the contributions are related to an ultra-low-power gunshot detector. Detailed discussion can be found in Chapter 2.

I/1. I proposed a novel acoustic delay line wake-up mechanism, implemented an experimental hardware, and showed that it can improve the power-consumption efficiency of audio event detectors.

I/2. I designed and implemented the hardware and software of an embedded gunshot detector module that utilizes the proposed wake-up mechanism and can be integrated into widely-used GPS tracking collars.

I/3. I developed a novel gunshot detector algorithm that employs the two-domain audio information used for the proposed wake-up mechanism, and evaluated its accuracy and efficiency through real-world experiments.

I/4. I developed a randomized architecture-search algorithm that generated, trained, and compared 1D and 2D convolutional neural networks that utilize the two-domain audio information for gunshot detection.

In the **second thesis group**, the contributions are related to the theoretical and practical investigations of the microphone mode of loudspeakers (referred to as reverse mode). Detailed discussion can be found in Chapter 3.

II/1. I proposed the idea of using loudspeakers for audio event detection by employing their reverse mode functionality. I carried out the theoretical modeling and analysis of the reverse mode, and supported it by real experiments.

II/2. I investigated through simulation experiments the reverse mode speakers in acoustic event detection scenarios.

II/3. I designed and implemented an audio event detector device based on the reverse mode functionality, and demonstrated its use by a simple, data-driven clap detector.

II/4. I investigated the loudspeakers' active reverse mode through theoretical modeling and analysis, and some experiments, when the speakers may be used for acoustic event detection while they are actively radiating sound.

In the **third thesis group**, the contributions are related to the automation of pupillometry and related image processing methods. Detailed discussion can be found in Chapter 4.

III/1. I developed and evaluated a pupil measurement method that is based on an energy attenuation model, implemented an automated feature extractor, and introduced new pupillogram features.

III/2. I redesigned the previously used pupillometry experimentation setup with a hardware extension on the camera, which enhanced the video quality, and thus supports more robust and more efficient experimentation.

III/3. I trained a fully-convolutional neural network for pupil segmentation that efficiently processes the videos acquired by the revised experimentation setup.

Összefoglalás

Az értkezés adatelemző alkalmazásokat ismertet, magába foglalva alacsony szintű hangfelismerő és magas szintű képfeldolgozó eljárásokat. A bemutatott megközelítések közös vonása, hogy hardver-közeli vagy hardvert érintő változtatások véghezvitele és a kapcsolódó szoftverek ezeket kihasználó együttes fejlesztése járult hozzá ahhoz, hogy az alkalmazások lehetségessé, egyszerűbbé vagy pontosabbá váltak.

A munka három fő témakörből áll. A 2. fejezetben egy állatok által hordozható lövésdetektort mutattam be, amelynek alkalmazhatóságát egy újszerű ébredési mechanizmussal tettem lehetővé. Ugyancsak hangfelismerés témában végzett kutatásaimat ismertettem a 3. fejezetben, amelyben megvizsgáltam a hangszórók mikrofonként való alkalmazhatóságát, amit egy egyszerű kiegészítő hardver tesz lehetővé. Az utolsó, 4. fejezetben egy videofeldolgozási alkalmazást mutattam be, amely automatizálta és javította patkányok pupillometriai vizsgálatait, ezzel támogatva egy szkizofréniával kapcsolatos orvosi kutatást.

Hordozható lövésdetektor rendszer

Az orvvadászat kiemelt helyen szerepel a biológiai sokféleséget romboló jelenségek listáján. Visszaszorítására a klasszikus bűnüldözés módszereit alkalmazzák, azonban, a vadvilág vadászata távoli, elhagyatott területeken történik, ahol nincs emberi jelenlét, így a támadások észlelése nehéz. Továbbá, a nagytestű állatok orvvadászata militarizált és sokszor globális bűnszervezetek által támogatott. A helyi vadvédelmi szervek túlterheltek és tehetetlenek a bűnüldözési eszközök hiánya és a nagy területeken élő vadállomány monitorozásának nehézsége miatt. Fontos tehát az olyan technológiák fejlesztése, amelyek elősegítik a távoli vadvédelmet. Napjainkban egy ígéretes irány az állatok által hordozható eszközök, mint például a GPS nyomkövető nyakörvek alkalmazása, amelyek lehetővé teszik az állatok valósidejű követését.

A 2. fejezetben egy állatok által hordozható akusztikus lövésdetektor rendszert mutattam be, amely kiegészeíti a jelenleg elterjedt GPS nyomkövető nyakörvek funkcionalitását. A két rendszer ötvözésével valósidejű lövésriasztások küldhetők helyinformációval is kiegészítve. A fő kihívások a többéves üzemidő, a lövedék lökés-

hullámának jó minőségű rögzítése és a hibás detektálások számának minimalizálása voltak. Egy újszerű akusztikus késleltető csatorna alkalmazásával az energiafelhasználás 88%-kal csökkent, ugyanakkor a detekciós pontosság releváns mértékben javult. A lövésdetektor modul hardvere és szoftvere úgy lett optimalizálva, hogy beépíthető legyen elefántokon használt GPS nyomkövető nyakörvekbe, és az első prototípus felhelyezésre is került egy vadonélő elefántra. A rendszer kiértékeléséhez kontrollált kísérleteket is végeztünk. Egy adatbázist gyűjtöttünk, amely tartalmazott természeti zajokat, ütődésből eredő hangokat és valós lövéseket is. A javasolt lövésdetektálási algoritmus jó eredményt ért el az adathalmazon, minden lövést detektált és egyetlen hamis riasztást sem generált. Adatvezérelt megközelítéseket is ismertettem a fejezetben, egy véletlenszerűsített architektúra-kereső algoritmust legenerált, betanított és összehasonlított 1D és 2D konvolúciós neurális hálózatokat, amelyek a kétféle hanginformációt használják lövésdetektálásra.

Hangszórók hangrögzítőkként való alkalmazása

A 3. fejezetben megvizsgáltam a hangszórók mikrofonszerű üzemmódját (ún. fordított üzemmód). Ebben az üzemmódban a hangszóró a felületére érkező hanghullámokat elektromos jellé alakítja, ezzel környezete megfigyelhetővé válik. Az javasolt ötlet a hangszórók alkalmazása volt akusztikus esemény detektálásra. Minimális hardver kiegészítéssel ez a funkció biztosítható és egy egyszerű beágyazott rendszeren implemetálható.. Ez a rendszer biztosítja a hangszóró rendeltetésének megfelelő hangsugárzó működést, illetve további képességgekkel is felruházza. Például, az így előálló hangszóró rendszer alkalmas lehet biztonsági rendszerekben való alkalmazásra, ahol a gyanús hangok felismerése kritikus feladat.

Munkámban elvégeztem a fordított üzemmód elméleti modellezését és elemzését, amit valós kísérletekkel, mérésekkel és szimulációkkal is alátámasztottam. További alkalmazási irányokat is bemutattam a fejezetben, illetve megterveztem és megvalósítottam egy, a hangszórók fordított üzemmódján alapuló beágyazott akusztikus eseménydetektor modult, amelynek használatát egy egyszerű, adatvezérelt tapsdetektorral is demonstráltam. Az eddig vázolt vizsgálatok mind feltételezték, hogy a hangrögzítés időszakában a hangszóró tétlen, vagyis a meghajtó egysége inaktív. A fejezetben azonban kitérek azon bonyolultabb ún. aktív fordított üzemmódra is, amelyben a hangszórók akusztikus esemény detektálásra is alkalmazhatók, miközben aktívan sugároznak.

Pupillometria automatizálása

Egy pupillometriai alkalmazás automatizálását mutattam be az értekezés 4. fejezetében. A kapcsolódó orvosi kutatás a szkizofrénia mentális betegség objektíven is mérhető, a központi idengrendszerre gyakorolt hatásait vizsgálja egy patkány modellben a pupilla-fényreflex elemzése által.

A hagyományos pupillometriai vizsgálatok során a pupilla fényre való reakcióját videófelvételen rögzítik, majd az egyes képkockákon a pupilla méretét meghatározzák, ezzel előáll az ún. pupillogram. Az orvosi munka támogatása és gyorsítása érdekében a pupilla-anottációs lépést egy automatizált eljárásra cseréltem, amely képes a pupilla átmérőjének mérésére. Ez a megoldás egy újszerű, energia-elnyelődés alapú sugárkövetési módszer használatával képes a pupilla körvonalát meghatározni. Kezeli a széles tartományba eső képintenzitás változásokat, az alacsony kontrasztkülönbséget, a mozgásból származó elmosódást és a pupilla régió részleges kitakarásait, ezzel az alacsony minőségű videók is feldolgozhatók. Az eljárás kiértékeléséhez 20 kézzel címkézett videót használtam, amelyeken a rendszer 2% relatív pupilla-átmérő hibát ért el.

A kifejlesztett módszer hozzájárult jelentős orvosi eredmények bemutatásához, amelyek ismertették a központi idegrendszer megváltozott működését és a pupillakontroll csökkenését a vizsgált patkány-modellben. Néhány technikai jellegű nehézség leküzdése érdekében a kísérletek során használt felszerelések újraszervezését és fejlesztését javasoltam, amely felállás már nagyobb robusztusságot biztosított az állatkísérletek rögzítéséhez. Egy harver-alapú kiegészítés a kamerán csökkentette a mozgásból származó elmosdódások hatását és növelte a képen a jel-zaj viszonyt, ezzel jobb minőségű videókat eredményezett és a pupilla szegmentálási feladatot lényegesen egyszerűsítette. Ezt a szegmentálási feladatot egy teljesen-konvolúciós neurális hálózattal oldottam meg, amelynek tanításához a nyilvánosan is elérhető pupilla-szegmentációs adathalmazunkat használtam. A tesztképeken az módszer 96%-os pontosságot ért el.

A disszertáció tézisei

Az **első téziscsoportban** a hozzájárulásaim egy ultra-alacsony fogyasztású lövésdetektorhoz kapcsolódnak. A részletes bemutatás a 2. fejezetben található.

I/1. Javasoltam egy újszerű akusztikus késleltető csatorna alapú ébredési mechanizmust, amelyhez elkészítettem egy kísérleti hardvert is, továbbá megmutattam, hogy használatával javítható a hangalapú eseménydetektorok energiahatékonysága.

I/2. Megterveztem és megvalósítottam egy, a javasolt ébredési mechanizmust alkalmazó beágyazott lövésdetektor-modul hardver- és szoftver-rendszerét, ami integrálható elterjedt GPS nyomkövető nyakörvekbe.

I/3. Kifejlesztettem egy újszerű lövésdetektor algoritmust, amely kihasználja a javasolt ébredési mechanizmusból származó kétféle hanginformációt, és valós kísérletekkel vizsgáltam a pontosságát és hatékonyságát.

I/4. Kifejlesztettem egy véletlenszerűsített architektúra-kereső algoritmust, amely legenerált, betanított és összehasonlított 1D és 2D konvolúciós neurális hálózatokat, amelyek a kétféle hanginformációt használják lövésdetektálásra.

A **második téziscsoport** a hangszórók mikrofonszerű módjának (ún. fordított üzemmód) elméleti analíziséhez és alkalmazási lehetőségeiknek vizsgálatához kapcsolódik. A részletes bemutatás a 3. fejezetben található.

II/1. Javasoltam hangszórók alkalmazását akusztikus eseménydetektálási feladatokra, kihasználva a fordított üzemmódjukat. Elvégeztem a fordított üzemmód elméleti modellezését és elemzését, amit valós kísérletekkel is alátámasztottam.

II/2. Megvizsgáltam szimulációs módszerekkel a fordított üzemmódú hangszórók alkalmazhatóságát akusztikus eseménydetektálási feladatokban.

II/3. Megterveztem és megvalósítottam egy, a hangszórók fordított üzemmódján alapuló beágyazott akusztikus eseménydetektor modult, amely használatát egy egyszerű, adatvezérelt tapsdetektorral is demonstráltam.

II/4. Vizsgáltam a hangszórók ún. aktív fordított üzemmódját elméleti modellezéssel és elemzéssel, valamint néhány kísérlettel, melyekben a hangszórók akusztikus esemény detektálásra is alkalmazhatók, miközben aktívan sugároznak.

A **harmadik téziscsoport** hozzájárulásai egy pupillometriai alkalmazás automatizálásához és az ehhez szükséges képfeldolgozó eljárásokhoz kapcsolódnak. Részletes bemutatás a 4. fejezetben található.

III/1. Kidolgoztam és kiértékeltem egy energia-elnyelődésen alapuló modellt használó algoritmust pupillaátmérő mérésére, implemetáltam egy automatizált jellemzőkinyerő eljárást, továbbá bevezettem újszerű pupillogram jellemzőket.

III/2. Újraterveztem a korábban használt pupillometriai kísérleti eljárásmódot egy, a kamerát érintő hardver kiegészítéssel, amely jobb minőségű videókat eredményezett, ezáltal támogatja a robusztusabb és hatékonyabb kísérletezést.

III/3. Betanítottam egy teljesen-konvolúciós neurális hálózatot pupilla szegmentálásra, amely hatékonyan feldolgozza az új kísérleti összeállítással rögzített videókat.

Publications

Journal publications

[1] **G. Kalmár**, A. Büki, G. Kékesi, G. Horváth, and L. G. Nyúl. Image Processing-based Automatic Pupillometry on Infrared Videos. *Acta Cybernetica*, 23(2), 599-613, 2017.

[2] A. Büki, **G. Kalmár**, G. Kékesi, G. Benedek, L. G. Nyúl, and G. Horváth. Impaired pupillary control in "schizophrenia-like" WISKET rats. *Autonomic Neuroscience*, vol. 213, 34-42, 2018.

[3] **G. Kalmár**, A. Büki, G. Kékesi, G. Horváth, and L. G. Nyúl. Automating, Analyzing and Improving Pupillometry with Machine Learning Algorithms. *Acta Cybernetica*, 24(2), 197-209, 2019.

[4] **G. Kalmár**. Analysis and Utilization of Reverse Mode Loudspeakers. *IEEE Access, vol. 8., 66270-66280*, 2020.

Full papers in conference proceedings

[5] **G. Kalmár**, G. Wittemyer, P. Völgyesi, H.B. Rasmussen, M. Maróti, Á. Lédeczi. Animal-Borne Anti-Poaching System. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '19)*, Association for Computing Machinery, 91-102, 2019.

[6] **G. Kalmár**. Investigation of Reverse Mode Loudspeaker Performance in Urban Sound Classification. *27th European Signal Processing Conference (EUSIPCO)*, 1-5, 2019.

[7] **G. Kalmár**. Smart Speaker: Suspicious Event Detection with Reverse Mode Speakers. *42nd International Conference on Telecommunications and Signal Processing (TSP)*, 509-512, 2019.

Further related publications

[8] **G. Kalmár**, A. Büki, G. Kékesi, G. Horváth, and L. G. Nyúl. Image processing based automatic pupillometry on infrared videos. In *The 10th Jubilee Conference of PhD Students in Computer Science (CSCS): Volume of extended abstracts.*, 2016.

[9] **G. Kalmár**, A. Büki, G. Kékesi, G. Horváth, and L. G. Nyúl. Feature extraction and classification for pupillary images of rats. In *The 11th Jubilee Conference of PhD Students in Computer Science (CSCS): Volume of extended abstracts.*, 2018.

[10] **Kalmár G.**, Büki A., Kékesi G., Nyúl L., and Horváth G.. Pupillametria automatizálása, vizsgálata és javítása gépi tanuló algoritmusokkal. *Képfeldolgozók és Alakfelismerők Társaságának 12. Országos Konferenciája*, 2019.

[11] **G. Kalmár**, G. Wittemyer, P. Völgyesi, H.B. Rasmussen, M. Maróti, and Á. Lédeczi. Animal-Borne Acoustic Gunshot Detector (poster). In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys '19)*, Association for Computing Machinery, 578-579, 2019.

[12] A. Büki, **G. Kalmár**, G. Kékesi, G. Benedek, L. G. Nyúl, and G. Horváth. Characterization of pupillary response in "schizophrenia-like" (WISKET) rats. *5th FENS Regional Meeting 2017*, 2017.

[13] Büki A., **Kalmár G.**, Kékesi G., Nyúl L., and Horváth G.. Autonóm idegrendszeri eltérések vizsgálata transzlációs modellben. *A Magyar Élettani Társaság, a Magyar Kísérletes és klinikai Farmakológiai Társaság és a Magyar Mikrocirkulációs és Vaszkuláris Biológiai Társaság közös Vándorgyűlése*, 2017.

[14] Büki A., **Kalmár G.**, Kékesi G., Nyúl L., and Horváth G.. A pupilla fényreflex nembeli különbségeinek vizsgálata patkányban. In *Magyar Élettani Társaság 2018. évi Vándorgyűlése : előadás és poszter absztraktok*, 2018.

Acknowledgments

First of all, I would like to thank my supervisor, László Nyúl, for directing my PhD studies. I would also like to thank my colleagues and friends who helped me to realize the results presented here and to enjoy the period of my studies. Last, but not least, I wish to thank my wife and family for their constant love and support.